NGS data analysis

Tomáš Hron
thron@hpst.cz

Autorizovaný distributor Agilent Technologies

Produkty
Služby
E-shop pro molekulární biologii

Kontakty
Servisní požadavek

O nás
Kariéra

Knihovna Agilent
Odkazy

Novinky

Semináře

Novinky a trendy Agilent Technologies 15. 10. 2020

Přehled HPST webinářů (aktuální i ze záznamu)

Setkání uživatelů platformy VarSome Clinical

Program přednášky

- Bioinformatická analýza
- Galaxy
- NGS
- Analysis Workflows
- Praktická část
  - KIR genotyping
  - RNAseq analysis

Program přednášky

- **Bioinformatická analýza**
- Galaxy
- NGS
- Analysis Workflows
- Praktická část
  - KIR genotyping
  - RNAseq analysis

Ten simple rules for providing effective bioinformatics research support

Judit Kumuthini, Michael Chimenti, Sven Nahnsen, Alexander Peltzer, Rebone Meraba, Ross McFadyen, Gordon Wells, Deanne Taylor, Mark Maienschein-Cline, Jian-Liang Li, Jyothi Thimmapuram, Radha Murthy-Karuturi, Lyndon Zass

Published: March 26, 2020 • https://doi.org/10.1371/journal.pcbi.1007531

Project development

Rule 1: Collaboratively design experiment

Rule 2: Manage scope and expectations

Rule 3: Define and ensure data management

Data collection and generation

Rule 4: Manage the traceability of data

Rule 5: Determine how and what metadata are reported

Rule 6: Coordinate data and internet security

Data analysis

Rule 7: Control data quality throughout the project lifecycle

Rule 8: Identify suitable computational tools for data analysis

Rule 9: Track, record, and confirm workflow changes

Rule 10: Repurpose the data

"Data science done well looks easy – and that's a big problem for data scientists."

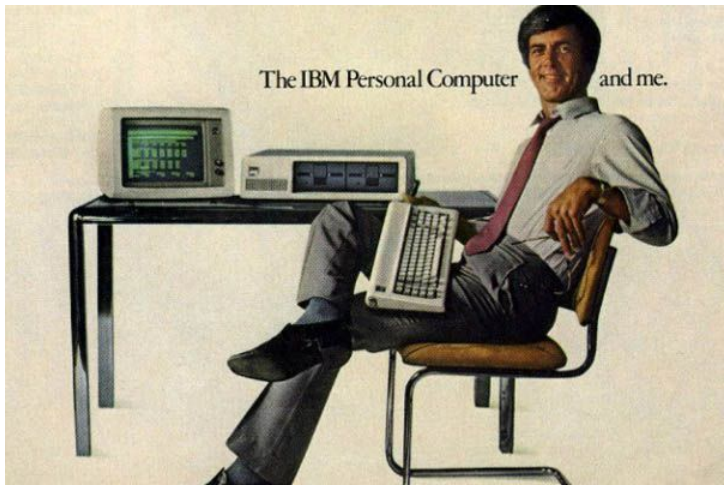
Jeff Leek

Associate Professor at Johns Hopkins Bloomberg School of Public Health.

Remote / cloud computing

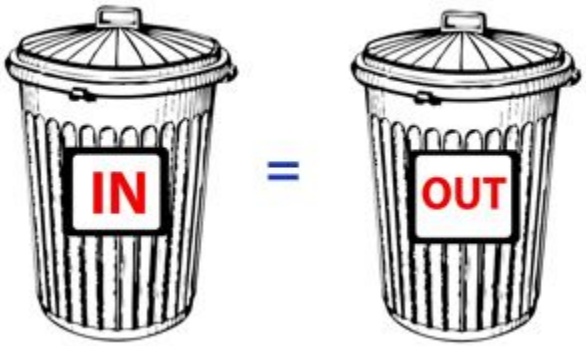
“There is no reason anyone would want a computer in their home.”

- Ken Olsen, 1977
Co-founder of Digital Equipment Corporation (DEC)



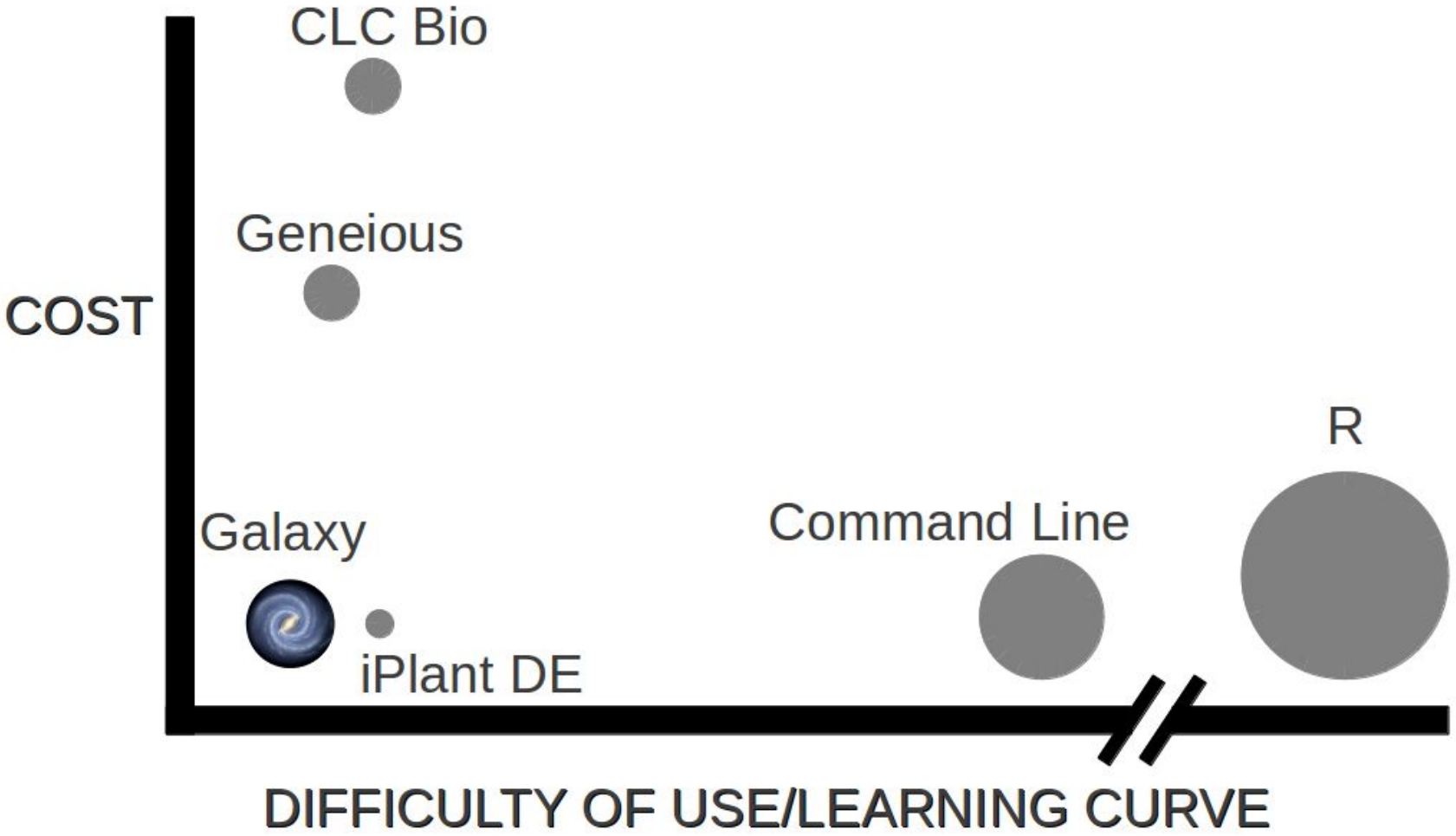
Dobrá data jsou předpokladem hladké analýzy a dobrého výsledku

Garbage in → Garbage out



Program přednášky

- Bioinformatická analýza
- **Galaxy**
- NGS
- Analysis Workflows
- Praktická část
  - KIR genotyping
  - RNAseq analysis



Size of dot indicates flexibility/power

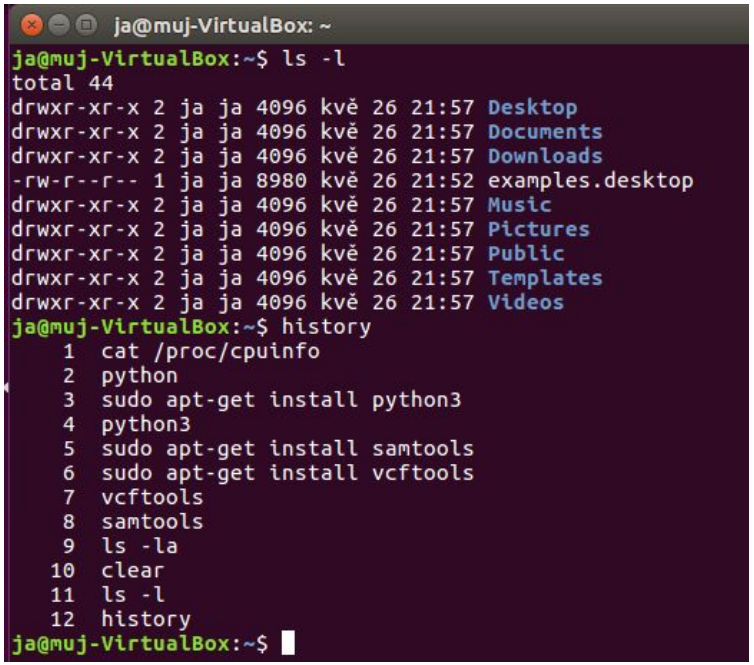
Co je Galaxy?

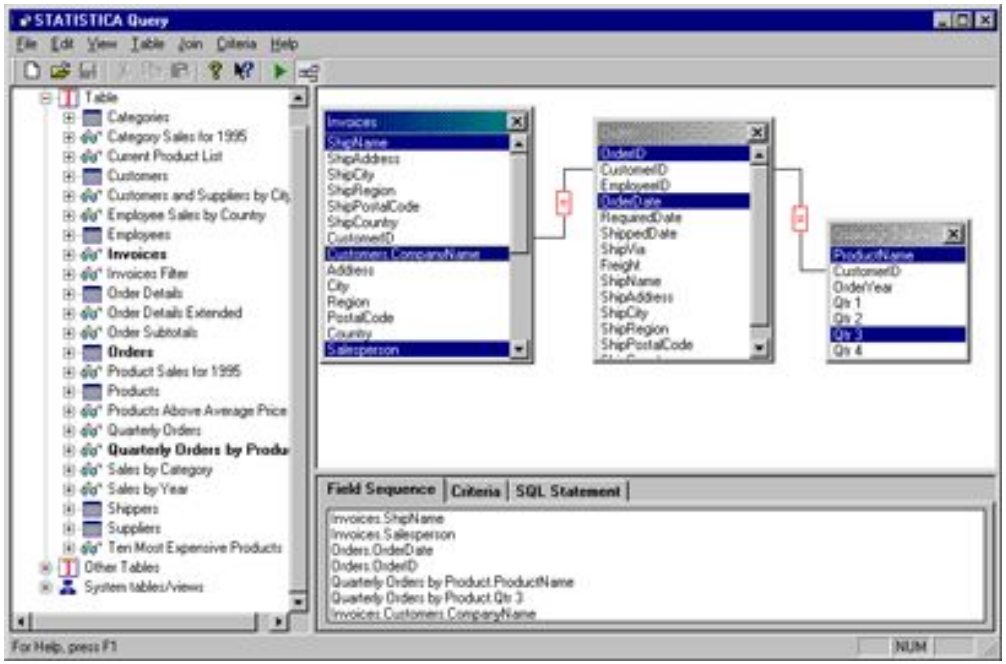
"Galaxy je otevřená webová platforma pro analýzu biomedicínských dat."

- Původně byla vytvořena pro genomiku, ale v současnosti obsahuje nástroje pro bioinformatickou analýzu obecně

Command Line Interface
vs
Graphical User Interface

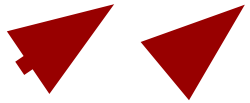
"Graphical user interfaces make easy tasks easy, while command line interfaces make difficult tasks possible"





odstranění nekvalitních čtení v Linuxu

```
thron@ntb-thron:~$
thron@ntb-thron:~$
thron@ntb-thron:~$
thron@ntb-thron:~$
thron@ntb-thron:~$
thron@ntb-thron:~$
thron@ntb-thron:~$
thron@ntb-thron:~$
thron@ntb-thron:~$
thron@ntb-thron:~$
thron@ntb-thron:~$ java -jar ~/Trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 6 -phred33 input_R1.fa
stq input_R2.fastq output_R1.fq output_U1.fq output_R2.fq output_U2.fq AVGQUAL:20
```



odstranění nekvalitních čtení v galaxy

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.3) Versions Options

Single-end or paired-end reads?

Single-end

Input FASTQ file

No fastqsanger or fastqsanger.gz dataset available.

Perform initial ILLUMINACLIP step?

Yes No

Cut adapter and other illumina-specific sequences from the read

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Drop reads with average quality lower than a specified level (AVGQUAL)

Minimum average quality required to keep a read

20

Insert Trimmomatic Operation

Execute

Jak dlouho existuje Galaxy?

Od roku 2005

Galaxy: A platform for interactive large-scale genome analysis

Belinda Giardine,[1] Cathy Riemer,[1] Ross C. Hardison,[1] Richard Burhans,[1] Laura Elnitski,[2] Prachi Shah,[1,2] Yi Zhang,[1] Daniel Blankenberg,[1] Istvan Albert,[1] James Taylor,[1] Webb Miller,[1] W. James Kent,[3] and Anton Nekrutenko[1,4]

[1]Center for Comparative Genomics and Bioinformatics, Huck Institutes for Life Sciences, Penn State University, University Park, Pennsylvania 16802, USA; [2]National Human Genome Research Institute, Bethesda, Maryland 20892, USA; [3]Department of Computer Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA

Accessing and analyzing the exponentially expanding genomic sequence and functional data pose a challenge for biomedical researchers. Here we describe an interactive system, Galaxy, that combines the power of existing genome annotation databases with a simple Web portal to enable users to search remote resources, combine data from independent queries, and visualize the results. The heart of Galaxy is a flexible history system that stores the queries from each user; performs operations such as intersections, unions, and subtractions; and links to other computational tools. Galaxy can be accessed at http://g2.bx.psu.edu.

[Supplemental material is available online at www.genome.org.]

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1240089/pdf/00151451.pdf

Kdo vytváří Galaxy?

PennState

Center for Comparative Genomics and Bioinformatics

JOHNS HOPKINS
UNIVERSITY

Department of Biology

A velké množství přispěvatelů z komunity Galaxy…

Jak začít používat Galaxy?

- Na veřejném serveru **usegalaxy.org**
- Pro akademiky a studenty na serveru Metacentra
- Instalace na serveru vaší laboratoře

Cesnet a MetaCentum

bezplatné využití výpočetní a úložné kapacity a řady aplikačních programů pro akademiky

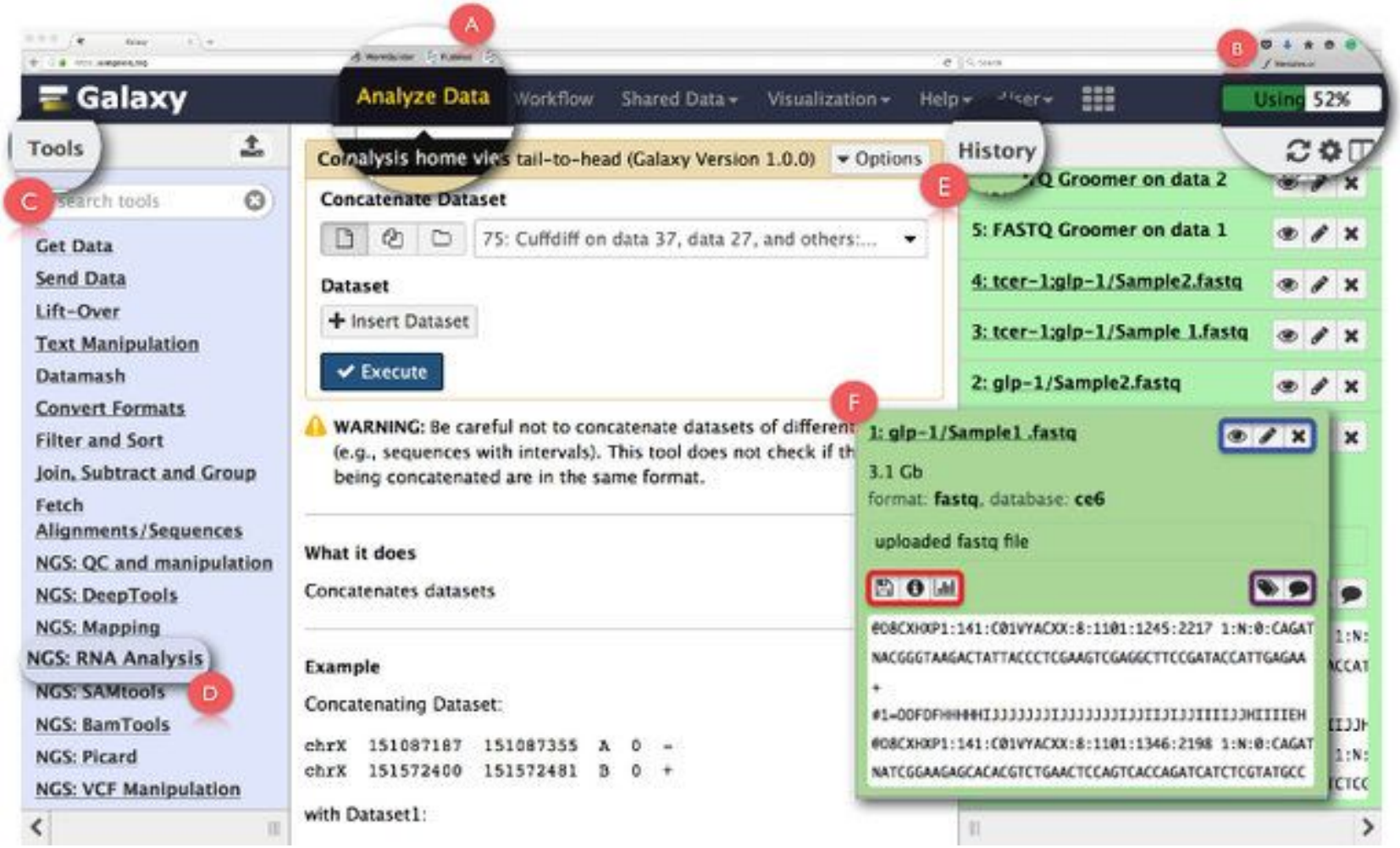
Důležité odkazy

Cesnet a MetaCentum

https://metavo.metacentrum.cz/ - MetaCentrum VO
https://wiki.metacentrum.cz/wiki/Main_Page - wiki stránka MetaCentra
https://wiki.metacentrum.cz/wiki/Galaxy_application - wiki Galaxy v MetaCentu
https://galaxy.metacentrum.cz/ - přístup na MetaCentrum Galaxy

Galaxy

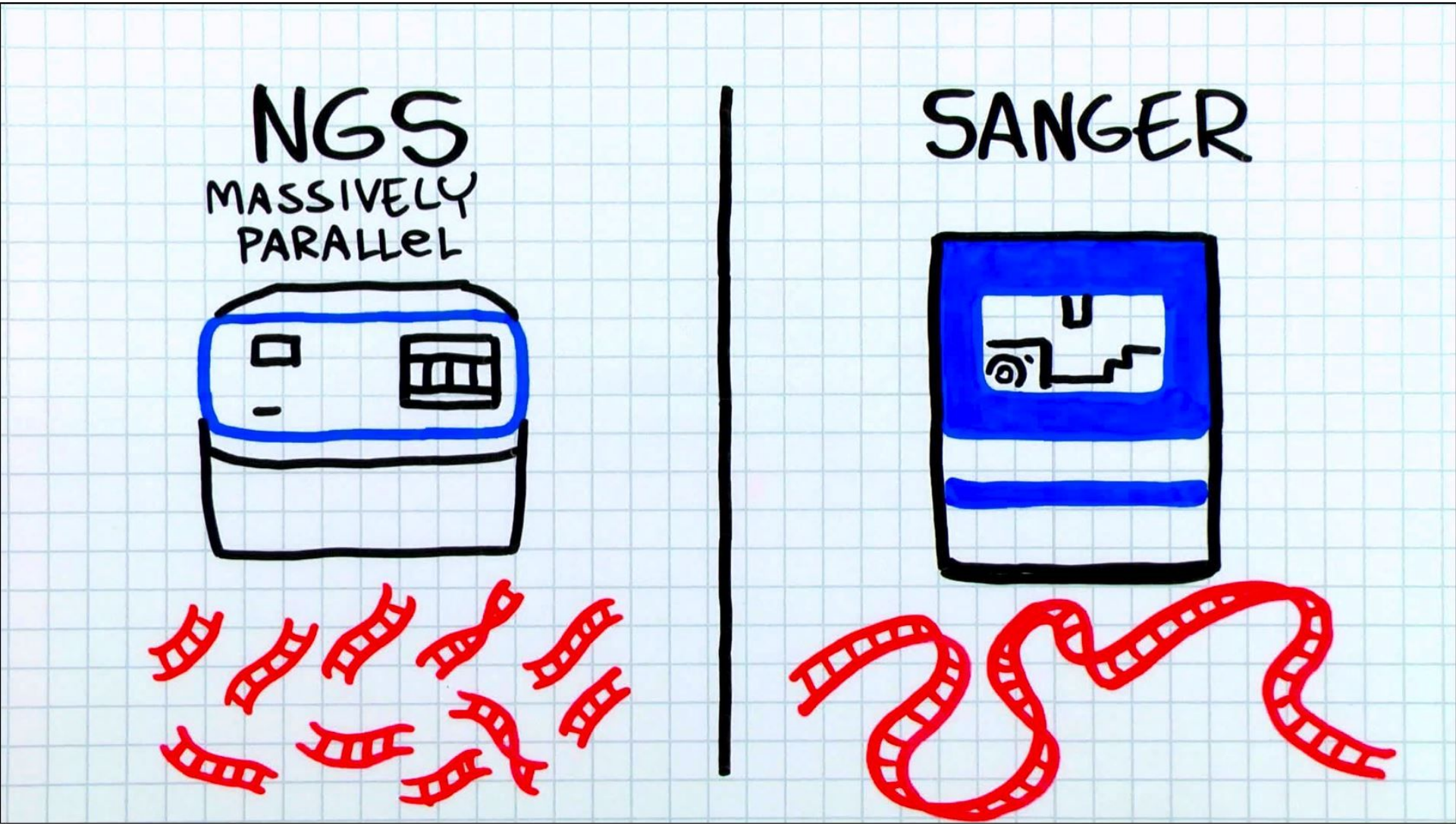
https://wiki.galaxyproject.org/Learn - Galaxy Wiki
https://vimeo.com/galaxyproject - Video tutoriály
http://wiki.bits.vib.be/index.php/Galaxy_beginner's_tutorial - Galaxy tutorial
https://www.coursera.org/learn/galaxy-project - Galaxy Coursera
https://biostar.usegalaxy.org/ - Galaxy Biostars



(A) highlights the 'Analyze data' function in the webpage header used to access Analysis Home View. (B) is the 'Progress bar' that indicates the space on the Galaxy server utilized by the operation. (C) is the 'Tools Section' that lists all the tools that can be run on the Galaxy interface. (D) shows the 'NGS: RNA Analysis' tool section used for RNA-Seq analysis. (E) depicts the 'History' panel that lists all the files generated using Galaxy. (F) shows an example of the dialogue box that opens up when clicking on any file in the History section. Within (F), the blue box highlights icons that can be used to view, edit the attributes or delete the dataset, the purple box highlights icons that can be used to 'edit' the dataset tags or annotation, and, the red box indicates icons to download the data, view details of the task performed or rerun the operation. Please click here to view a larger version of this figure.

Program přednášky

- Bioinformatická analýza
- Galaxy
- **NGS**
- Analysis Workflows
- Praktická část
  - KIR genotyping
  - RNAseq analysis


NGS
MASSIVELY
PARALLEL
SANGER

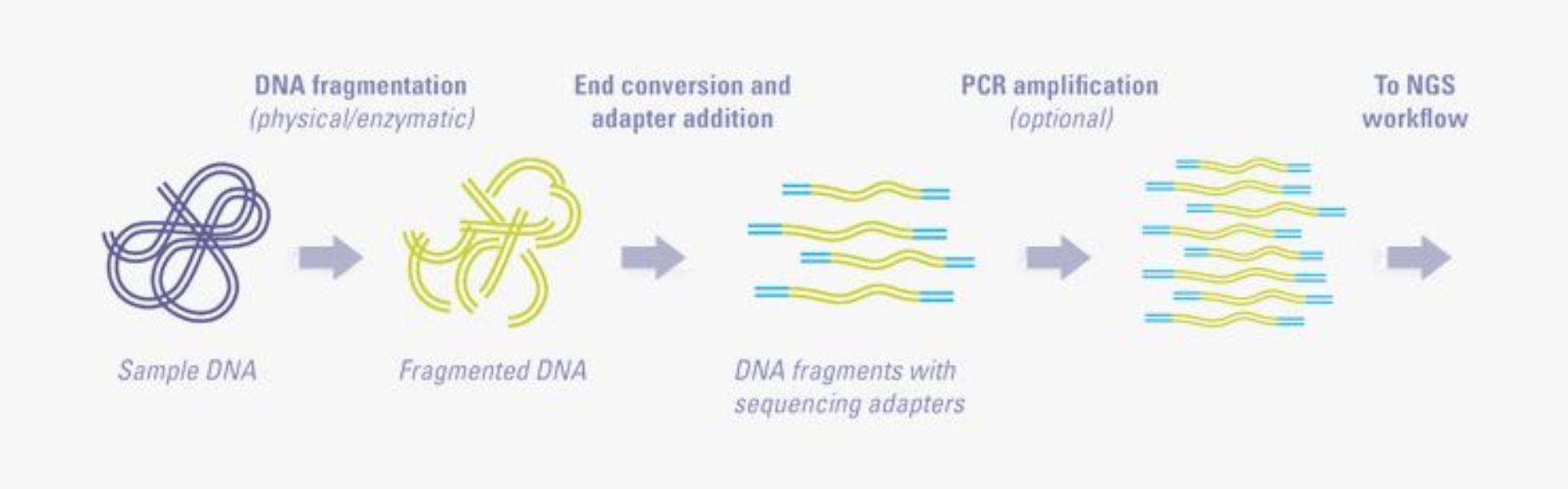
NGS (Next Generation Sequencing)

- Velké množství dat: od několika Gb po Tb
  - Vyžaduje automatizované zpracování dat
  - Vyžaduje infrastrukturu pro uskladnění dat
- Pokrytí: od několika genů po celý genom/transkriptom
- Vysoká citlivost ale náročnější interpretace pro klinické použití
- Volba velkého množství nástrojů pro zpracování dat

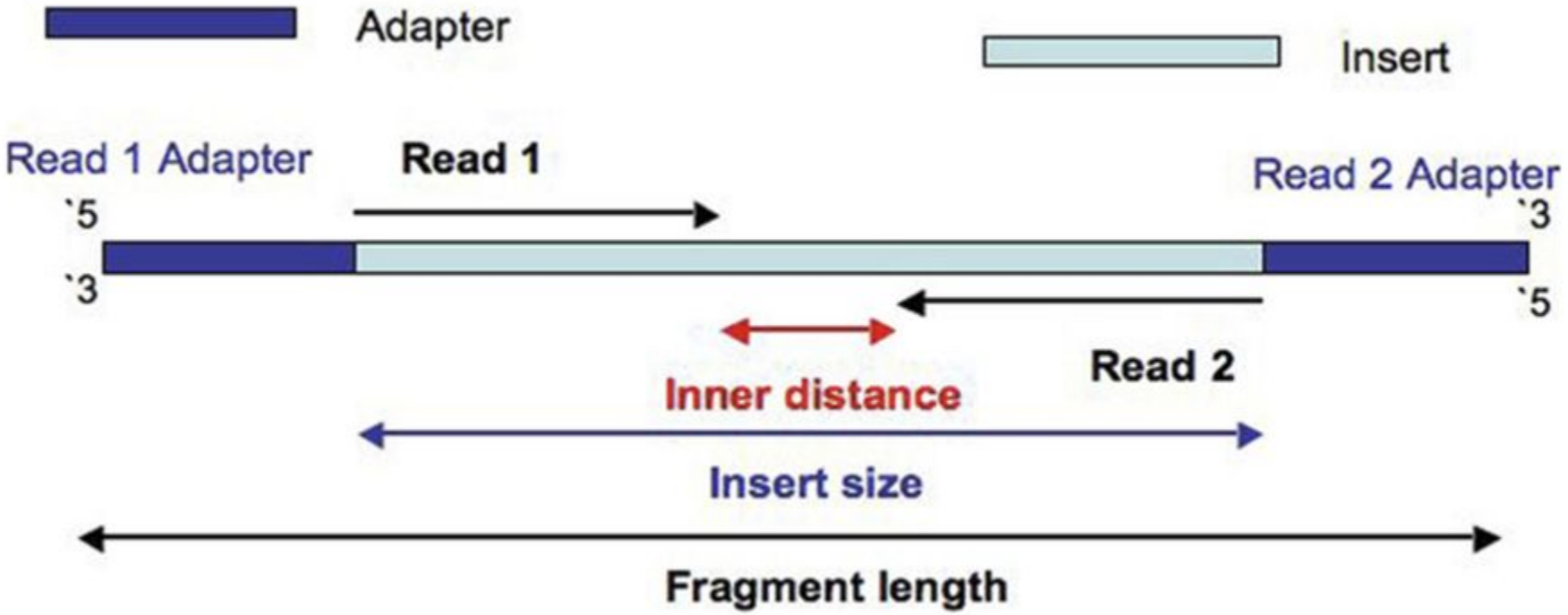
Illumina

- 4M - 400M reads per run
- Small read size (50 - 600bp)
- Accuracy 0.1 - 1%
- Whole genome sequencing or target enrichment

Sequencing libraries

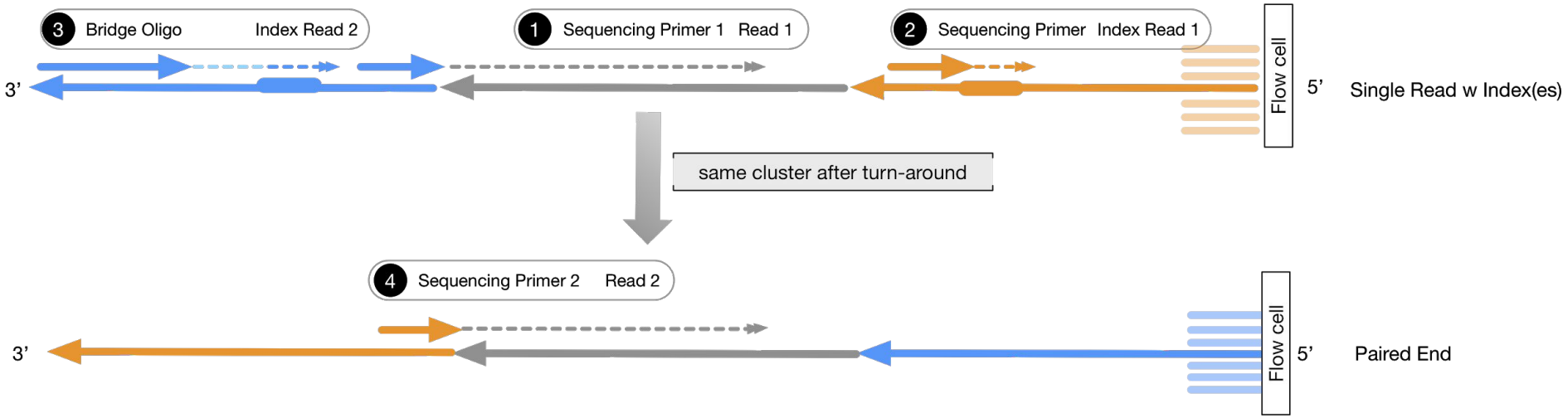


Sequencing libraries



Sequencing libraries

Indexes - scheme of sequencing procedure



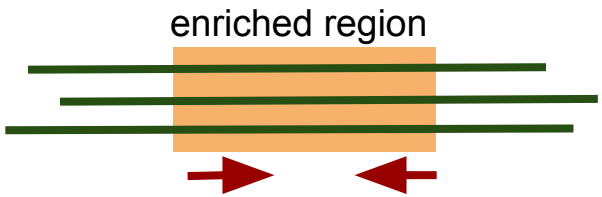
DNA sequencing libraries

Whole-genome lib.

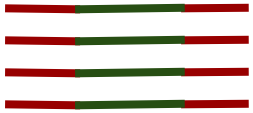
fragmentation of DNA

Sequencing

Amplicon lib.

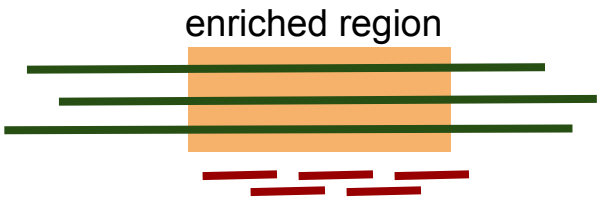


enrichment by specific primers

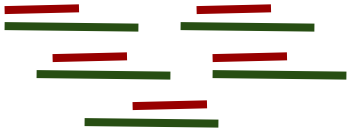


Sequencing

Sequence capture lib.

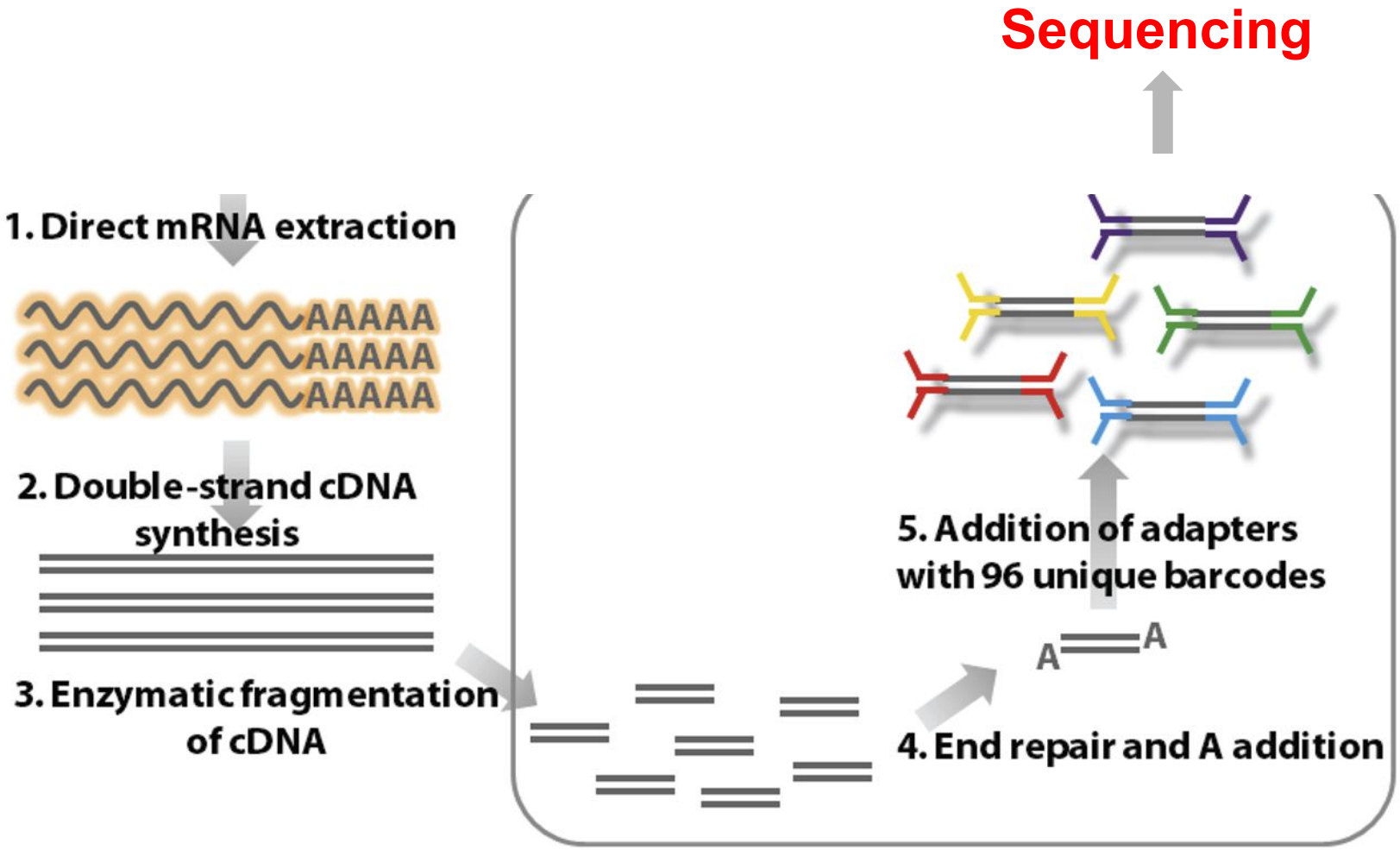


enrichment by specific probes



Sequencing

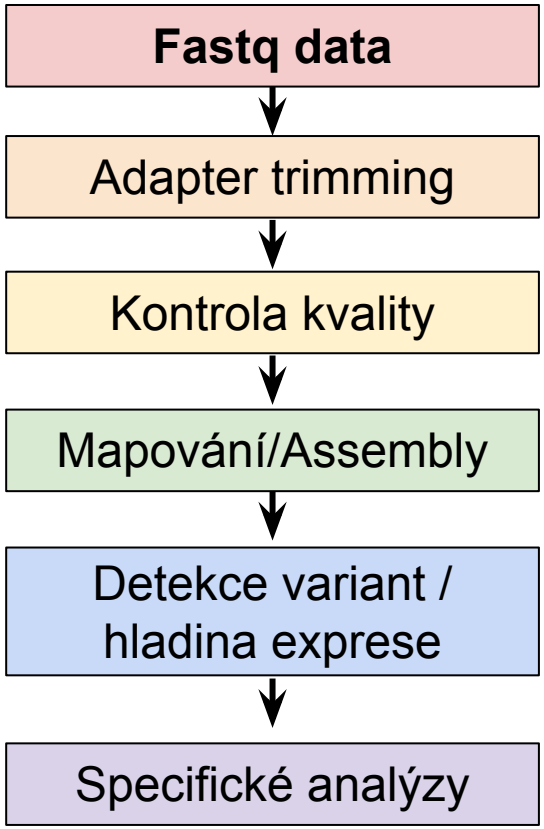
RNA sequencing libraries



Program přednášky

- Bioinformatická analýza
- Galaxy
- NGS
- **Analysis Workflows**
- Praktická část
  - KIR genotyping
  - RNAseq analysis

Analysis workflows



Analysis workflows

Fastq data

↓

Adapter trimming

↓

Kontrola kvality

↓

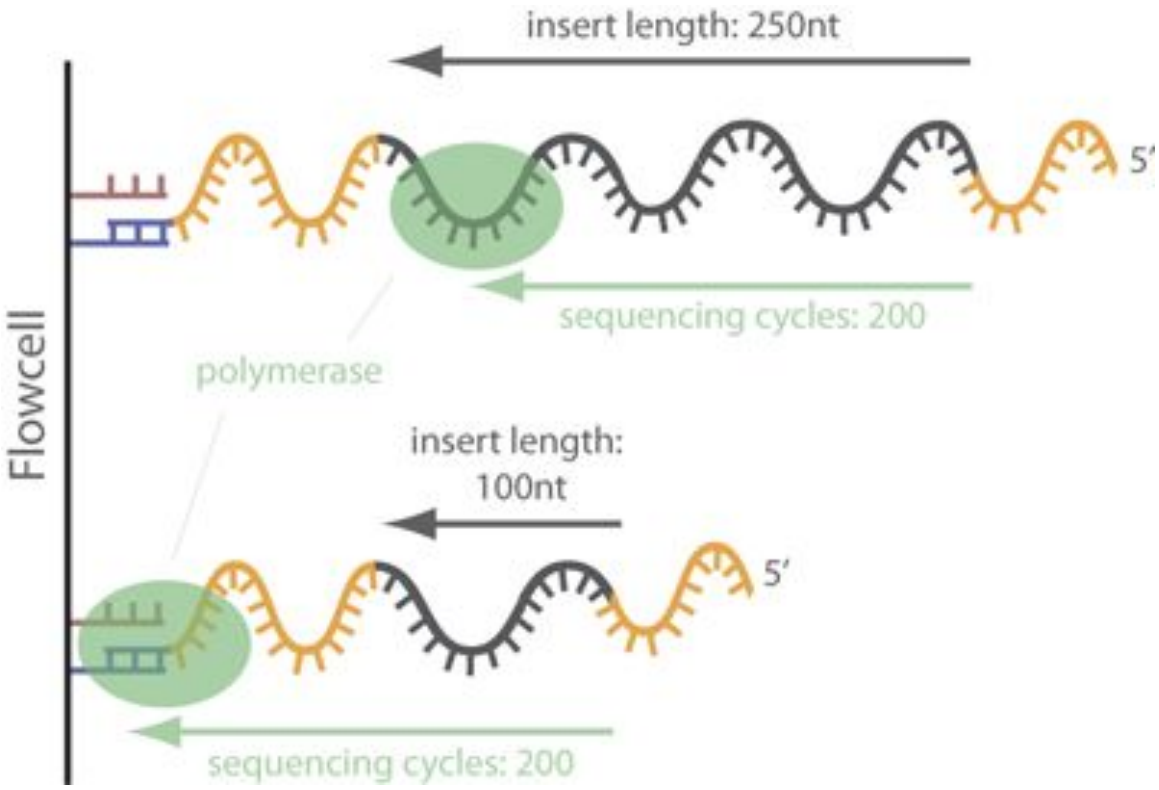
Mapování/Assembly

↓

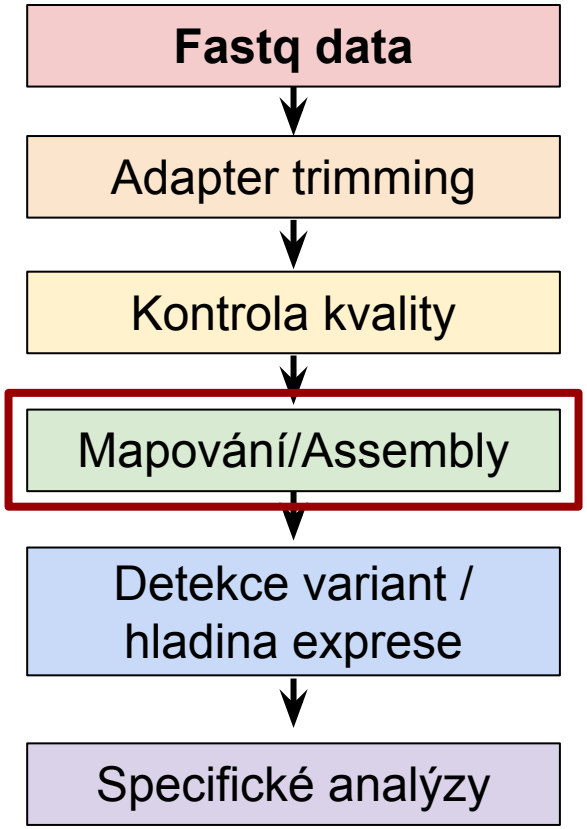
Detekce variant / hladina exprese

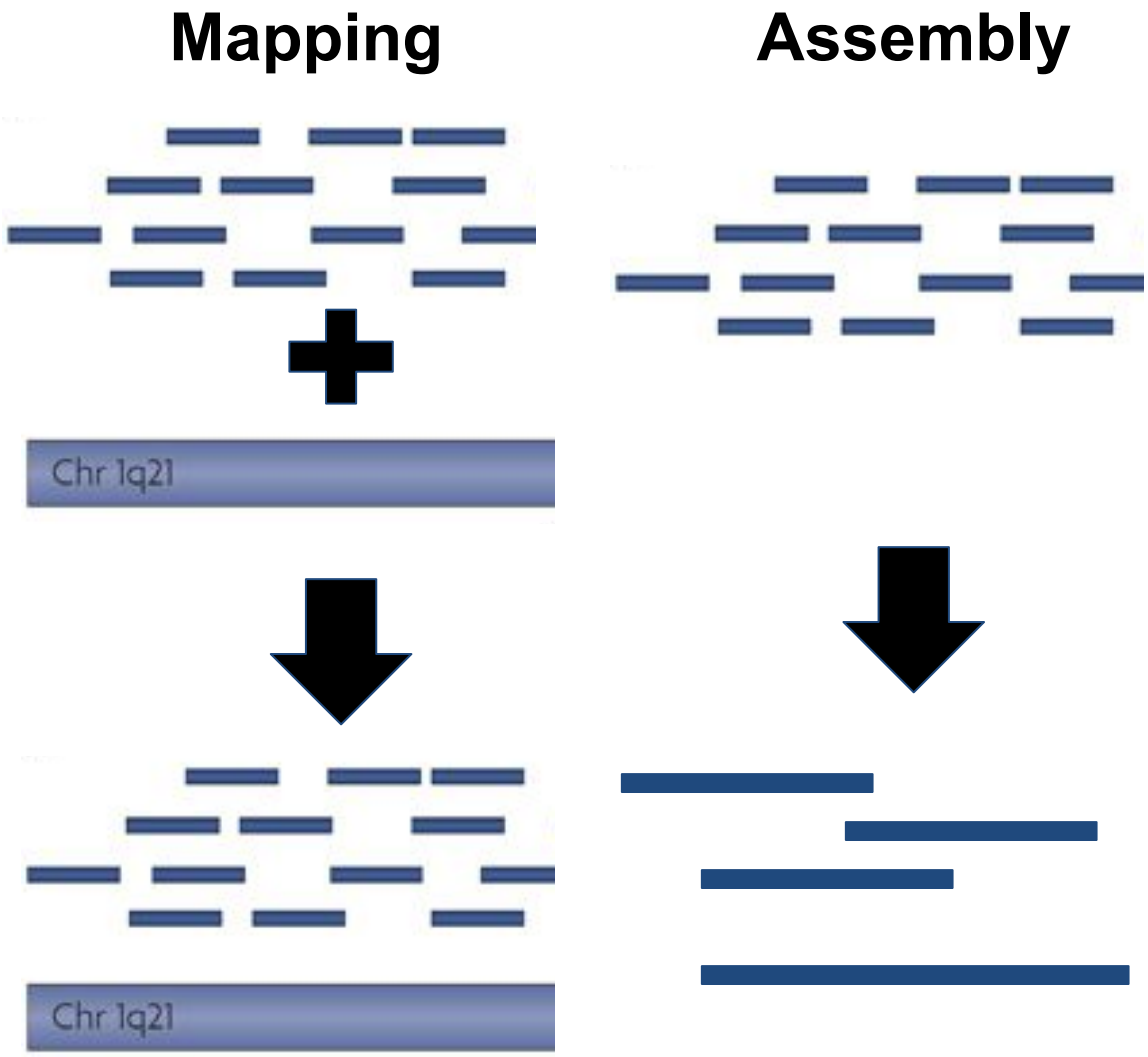
↓

Specifické analýzy

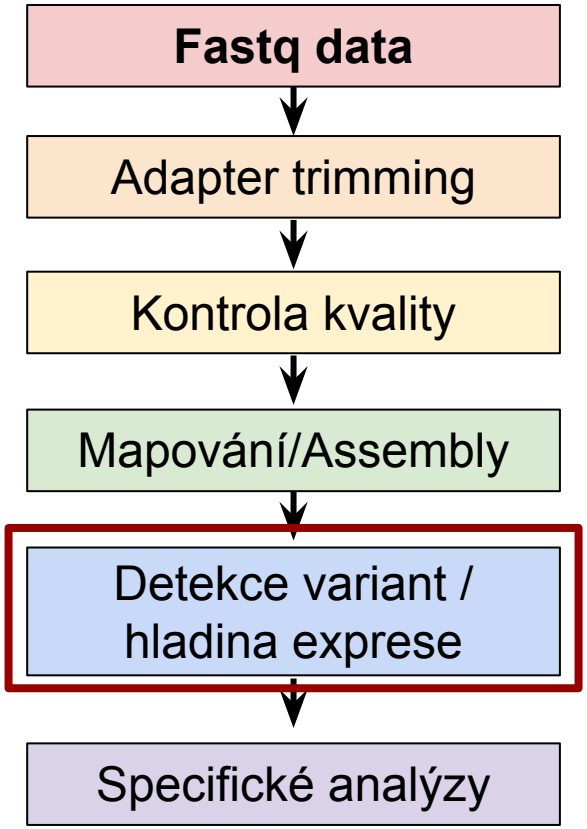


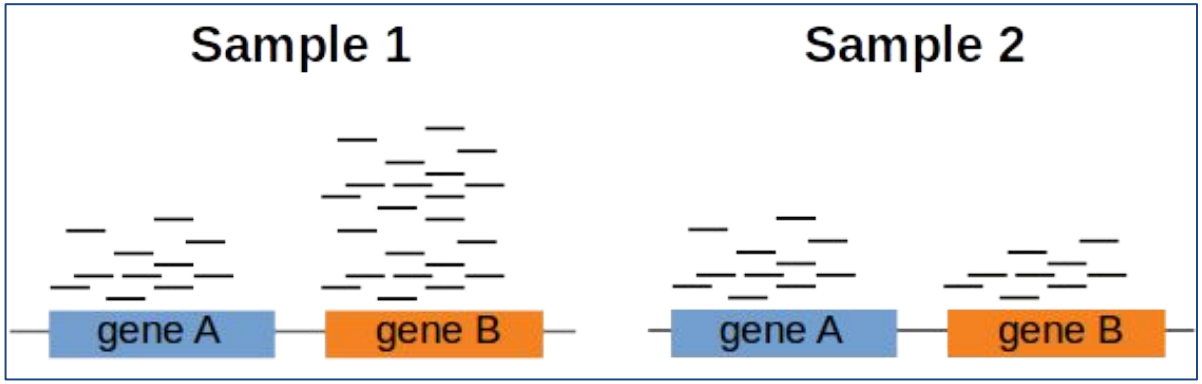
Analysis workflows



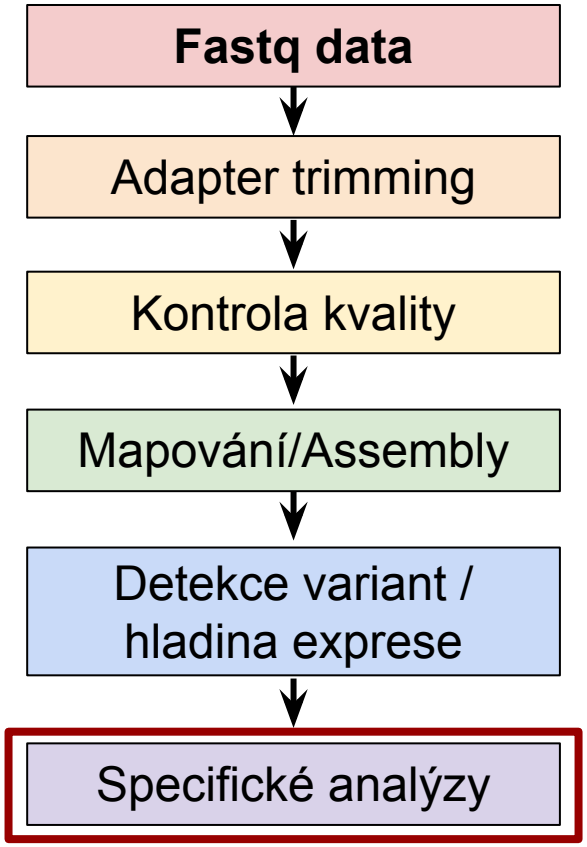


Analysis workflows





Analysis workflows



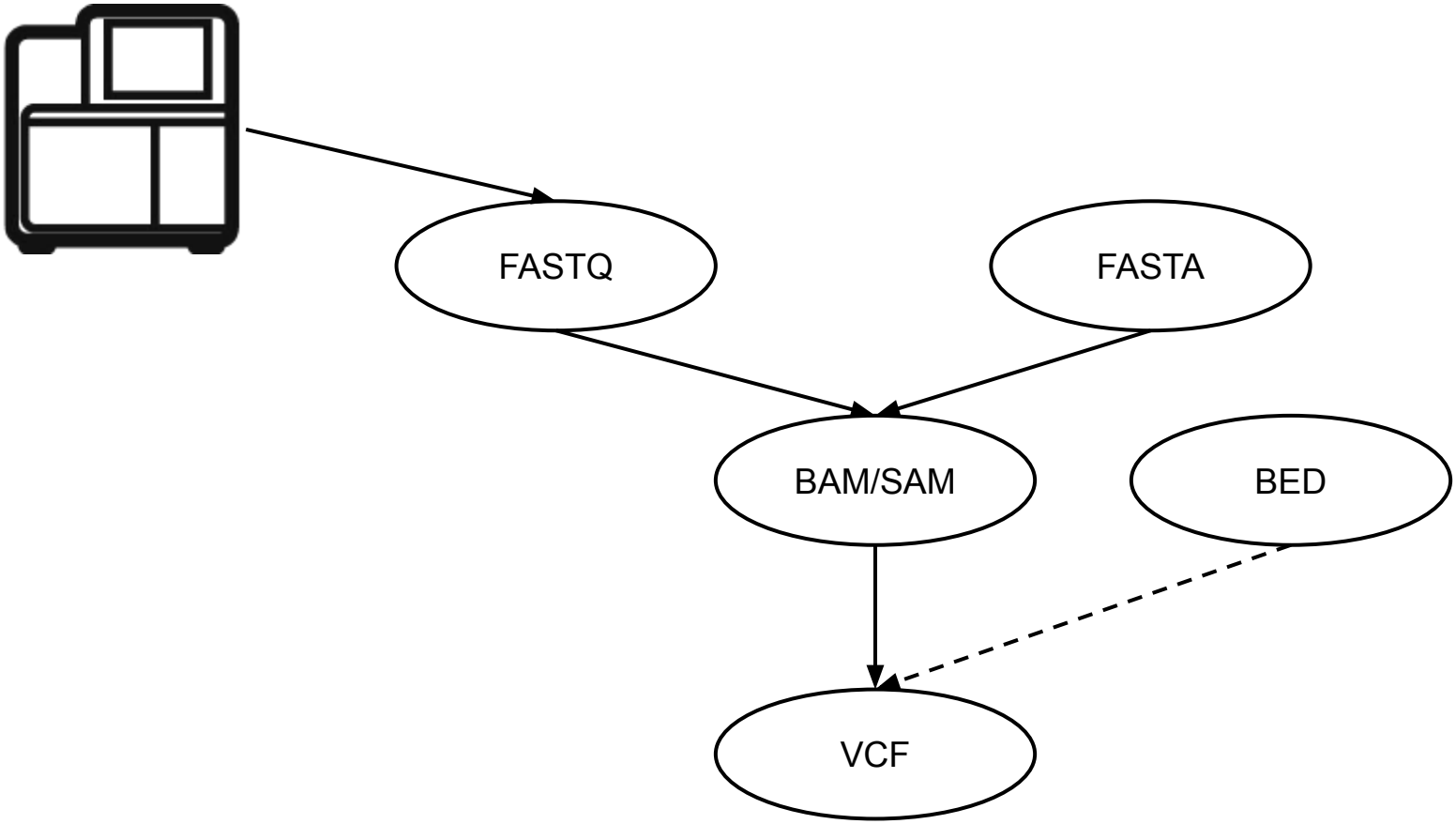
- **anotace variant**
- **efekt SNPs na kódující sekvence**
- **analýza diferenciální exprese**
- **de novo predikce miRNA**
- **identifikace mRNA isoforem**

.

.

.

Analysis workflows



Analysis workflows

Sekvence bazí
+ kvalita bazí

FASTQ

Sekvence bazí

FASTA

BAM/SAM

Sekvence bazí
+ jejich pozice
v ref genomu

BED

Seznam oblastí
v ref genomu

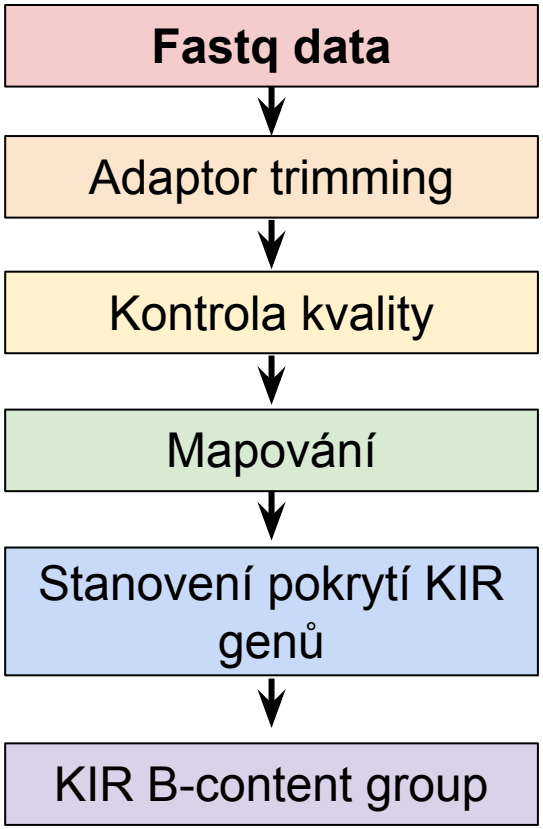
VCF

Seznam
detekovaných variant

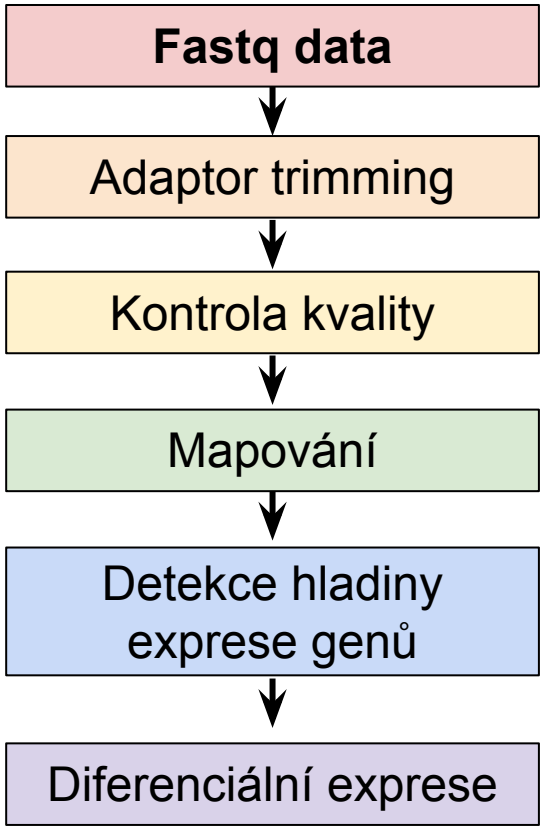
Program přednášky

- Bioinformatická analýza
- Galaxy
- NGS
- Analysis Workflows
- **Praktická část**
  - **KIR genotyping**
  - **RNAseq analysis**

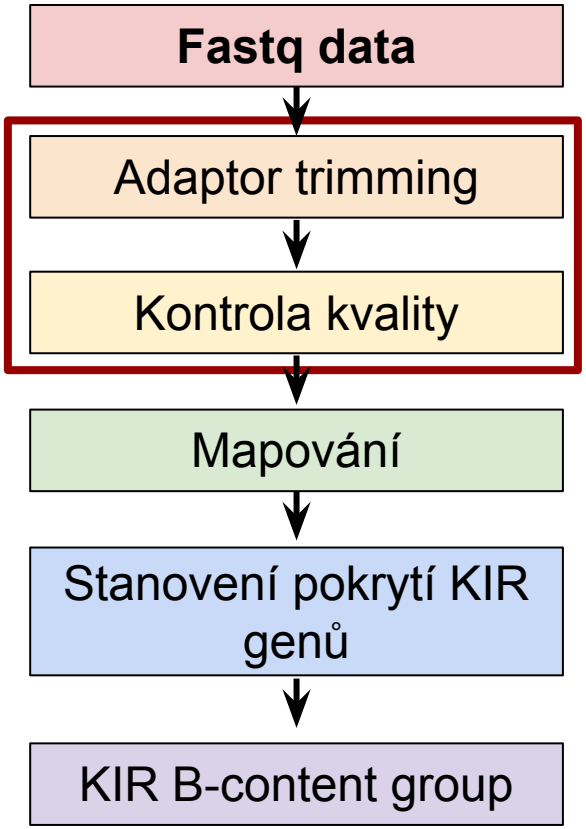
Genotypizace KIRs



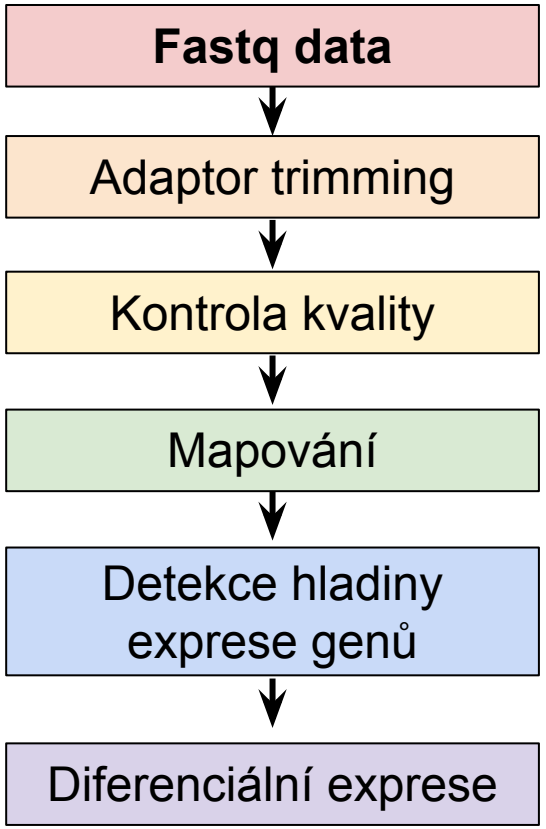
Diferenciální exprese genů u D. melanogaster



Genotypizace KIRs



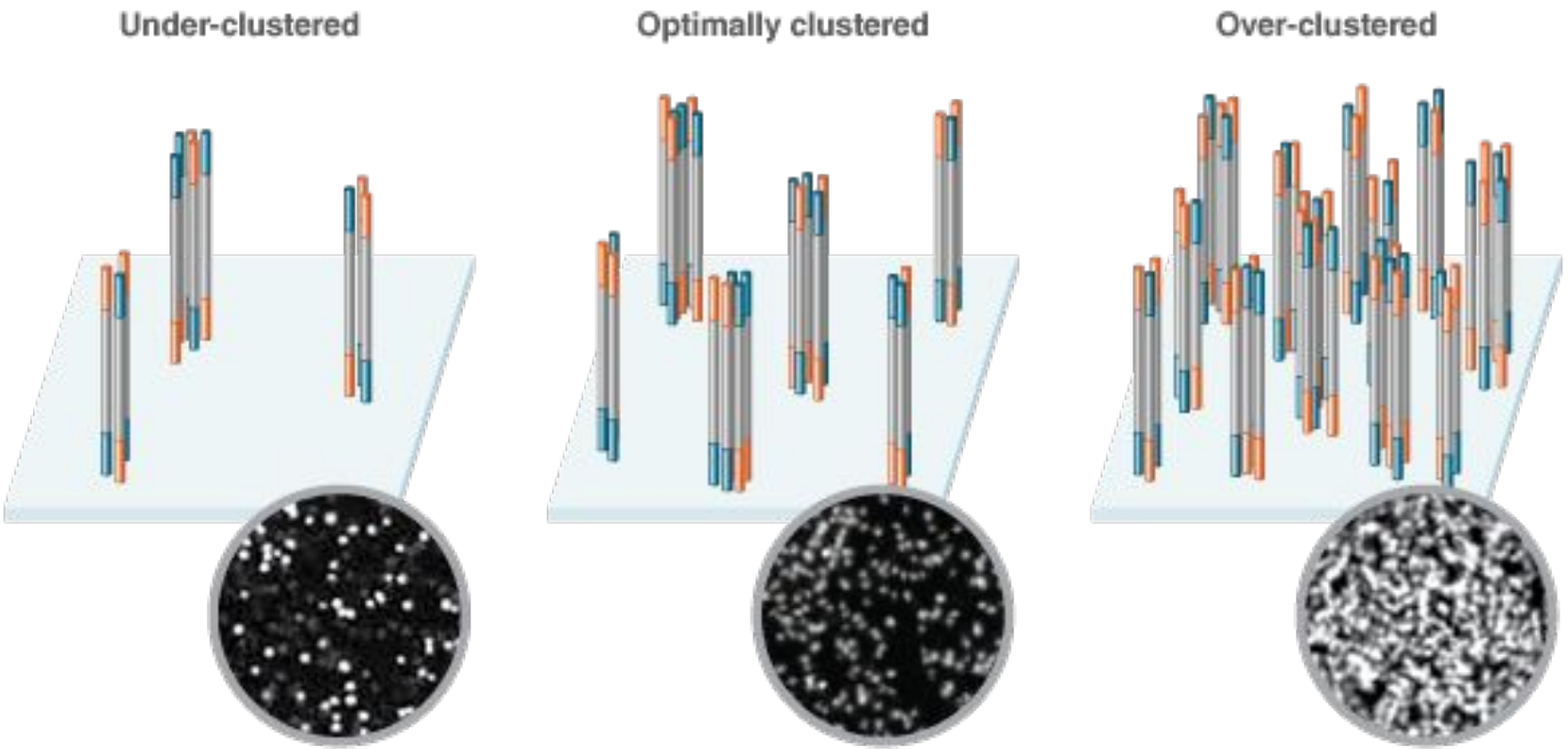
Diferenciální exprese genů u D. melanogaster



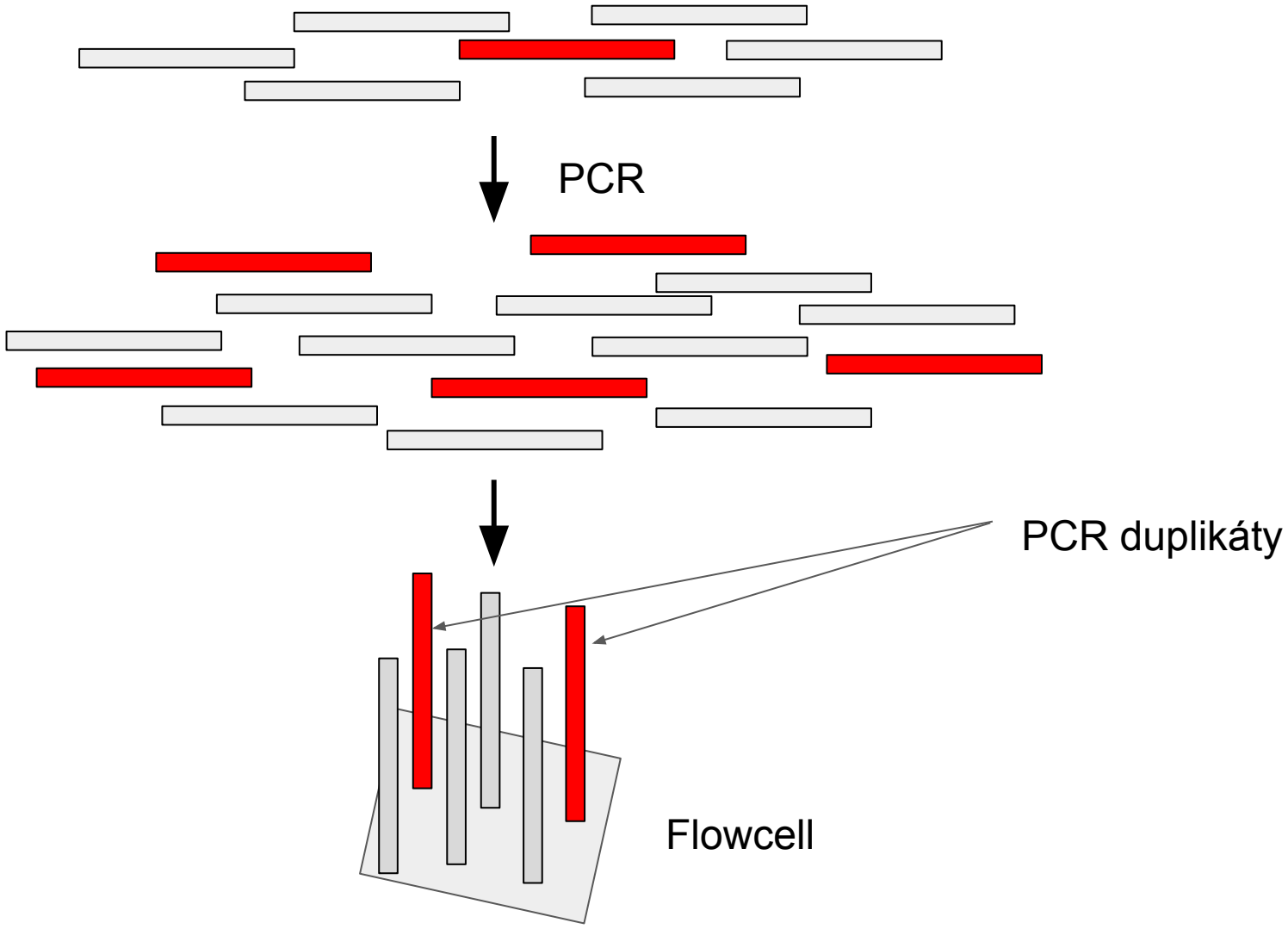
Illumina seq - co může být špatně?

- Špatné rozpoznání clusterů (overclustering)
- Přítomnost duplikátů
- Kontaminace adaptory
- Asynchronizace molekul clusteru (prephasing a postphasing)
- Plno dalších věcí...

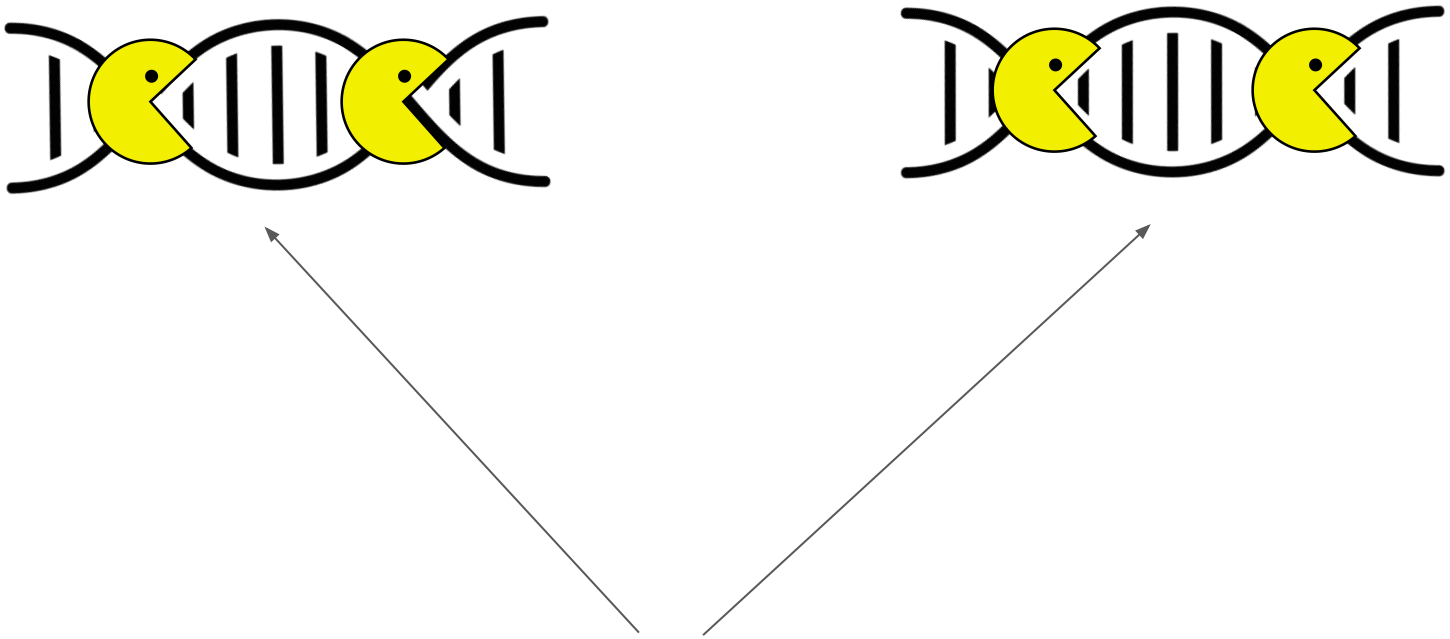
- Špatné rozpoznání svazků (overclustering)



PCR duplikáty



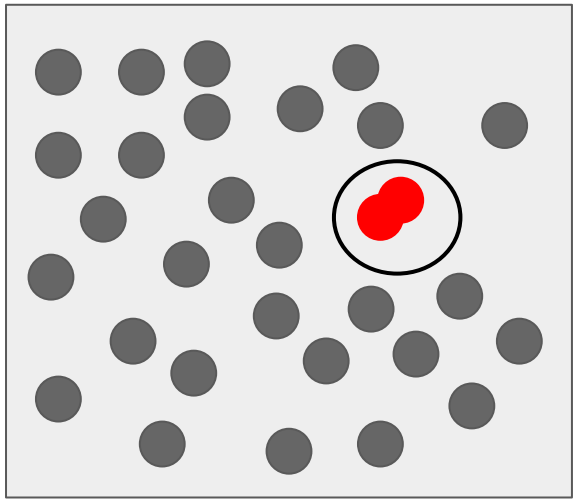
Štěpné duplikáty

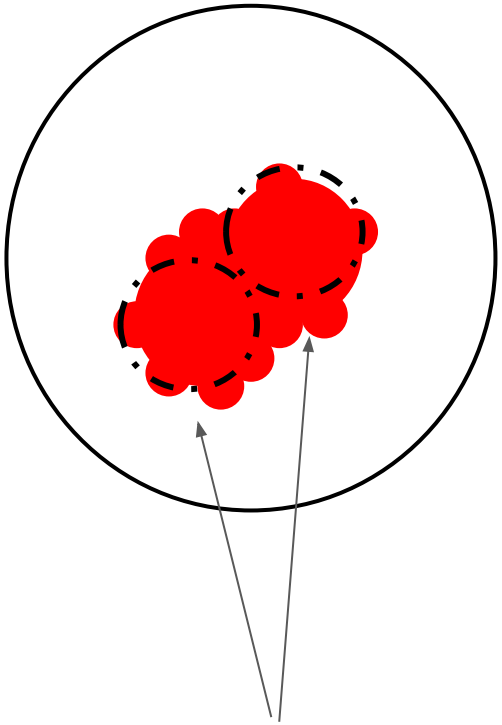


Štěpné duplikáty
(nenáhodné štěpení)

Optické duplikáty

Flowcell





1 svazek DNA molekul je brán kamerou jako 2 svazky

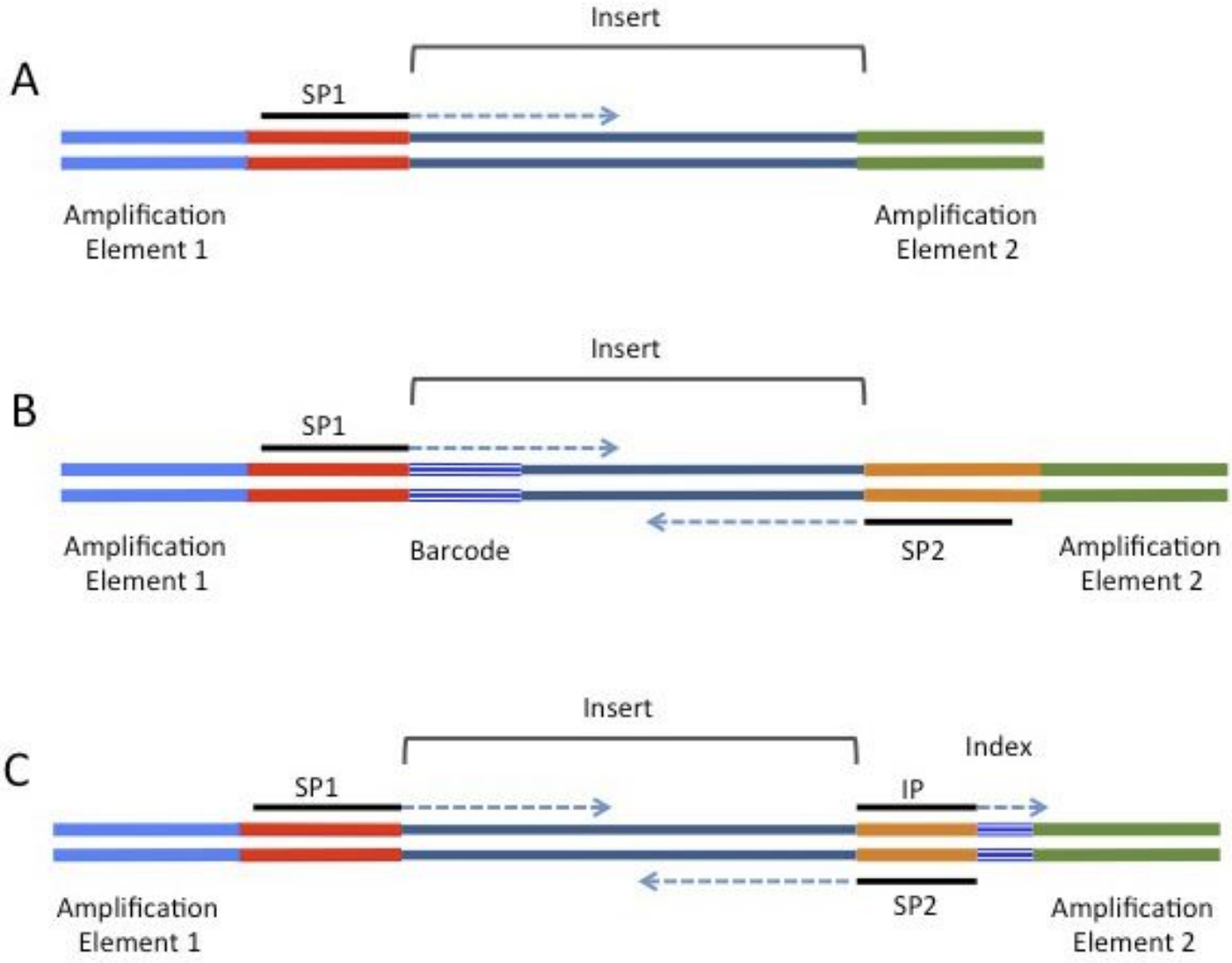
Přítomnost duplikátů

- **PCR duplikáty**: stejné molekuly DNA vzniklé při PCR se dostanou na různá místa flowcell
- **Optické duplikáty:** kamera zaznamená signál ze stejného místa na flowcell několikrát
- **Štěpné duplikáty:** stejné molekuly vznikají během nenáhodného štěpení DNA

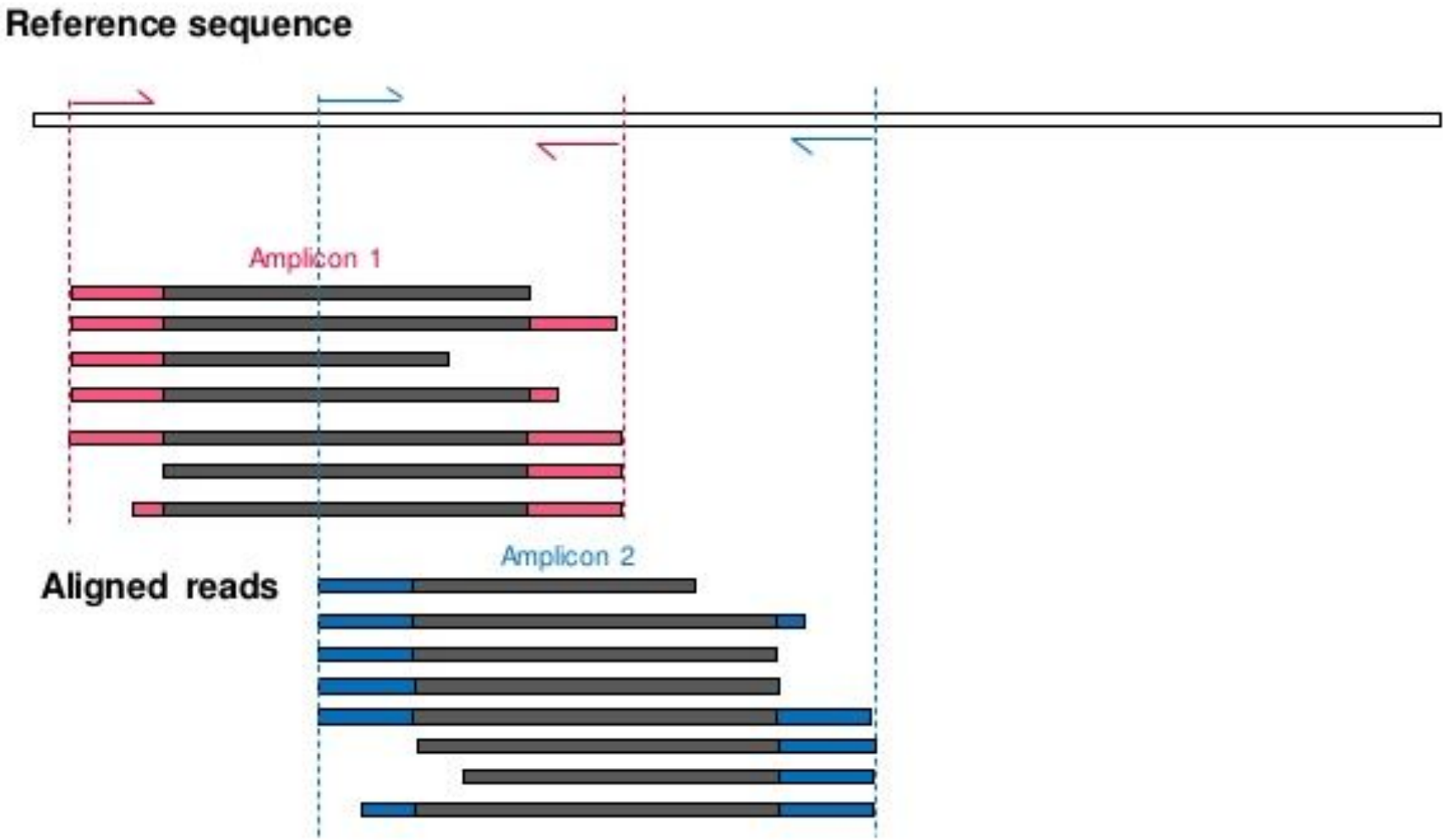
Přítomnost duplikátů

- Duplikáty mohou způsobit zkreslení výsledků, např. změna frekvence alel
- Správně identifikované PCR duplikáty pomáhají opravit chyby PCR a sekvenace

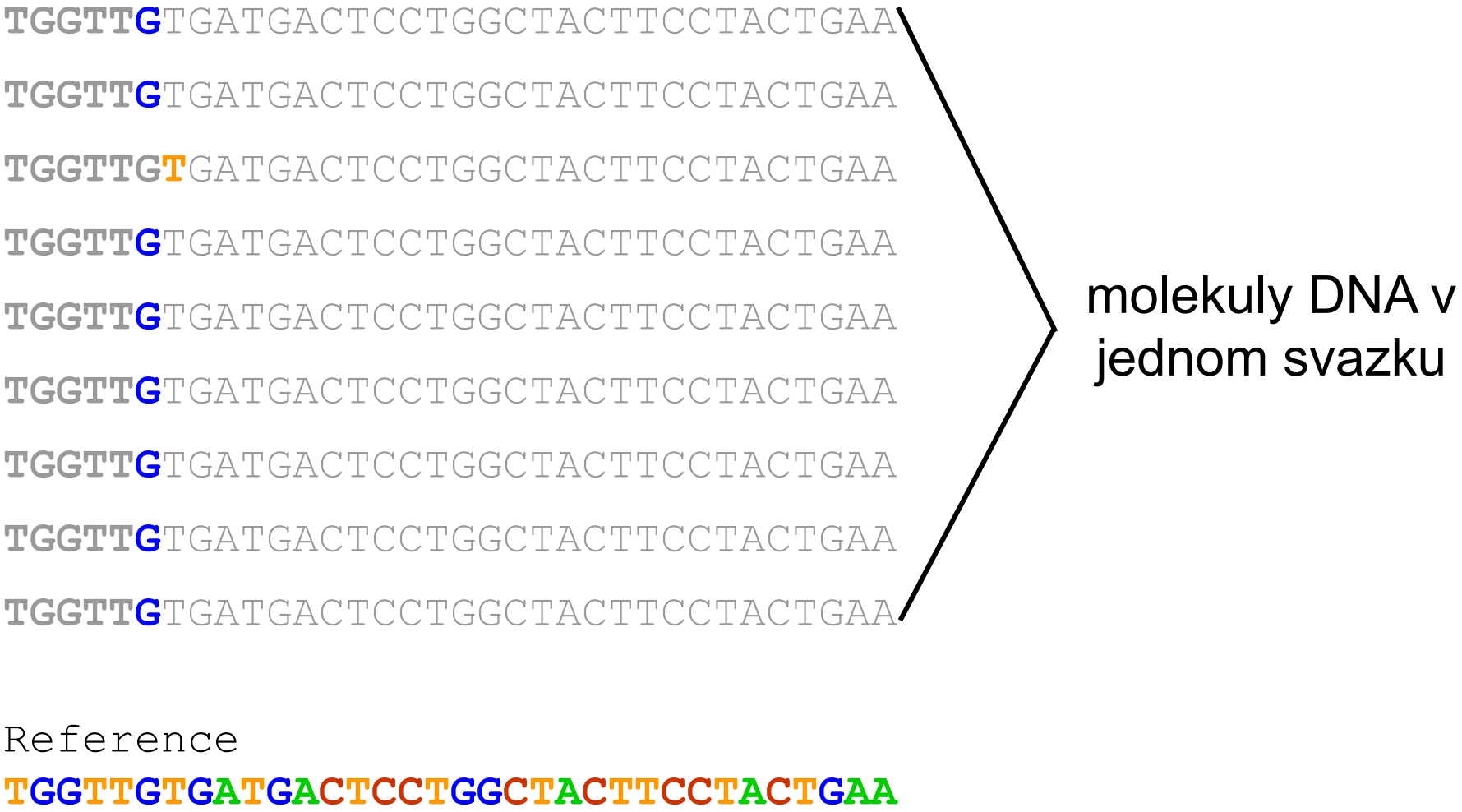
Kontaminace adaptory



Primery u amplikonových knihoven



- Asynchronizace molekul ve svazku (prephasing a postphasing)



- Asynchronizace molekul ve svazku (prephasing a postphasing)

```
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
```

Prephasing

```
Reference
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA
```

- Asynchronizace molekul ve svazku (prephasing a postphasing)

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA ⬅ Postphasing

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

Reference

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

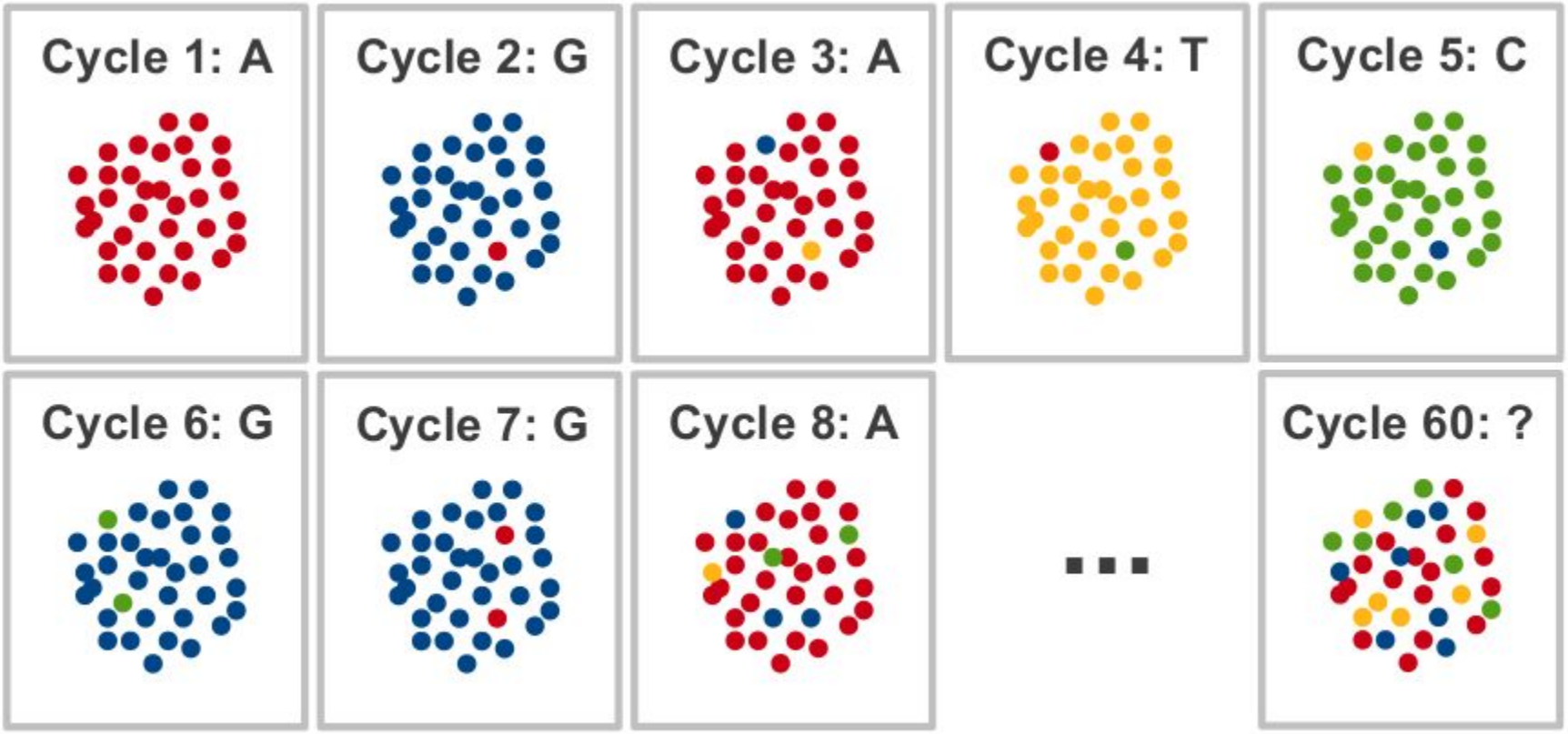
- Asynchronizace molekul ve svazku (prephasing a postphasing)

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA	**OK**
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA	-3
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA	+1
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA	**OK**
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA	**OK**
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA	-5
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA	+1
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA	+3
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA	**OK**

Reference

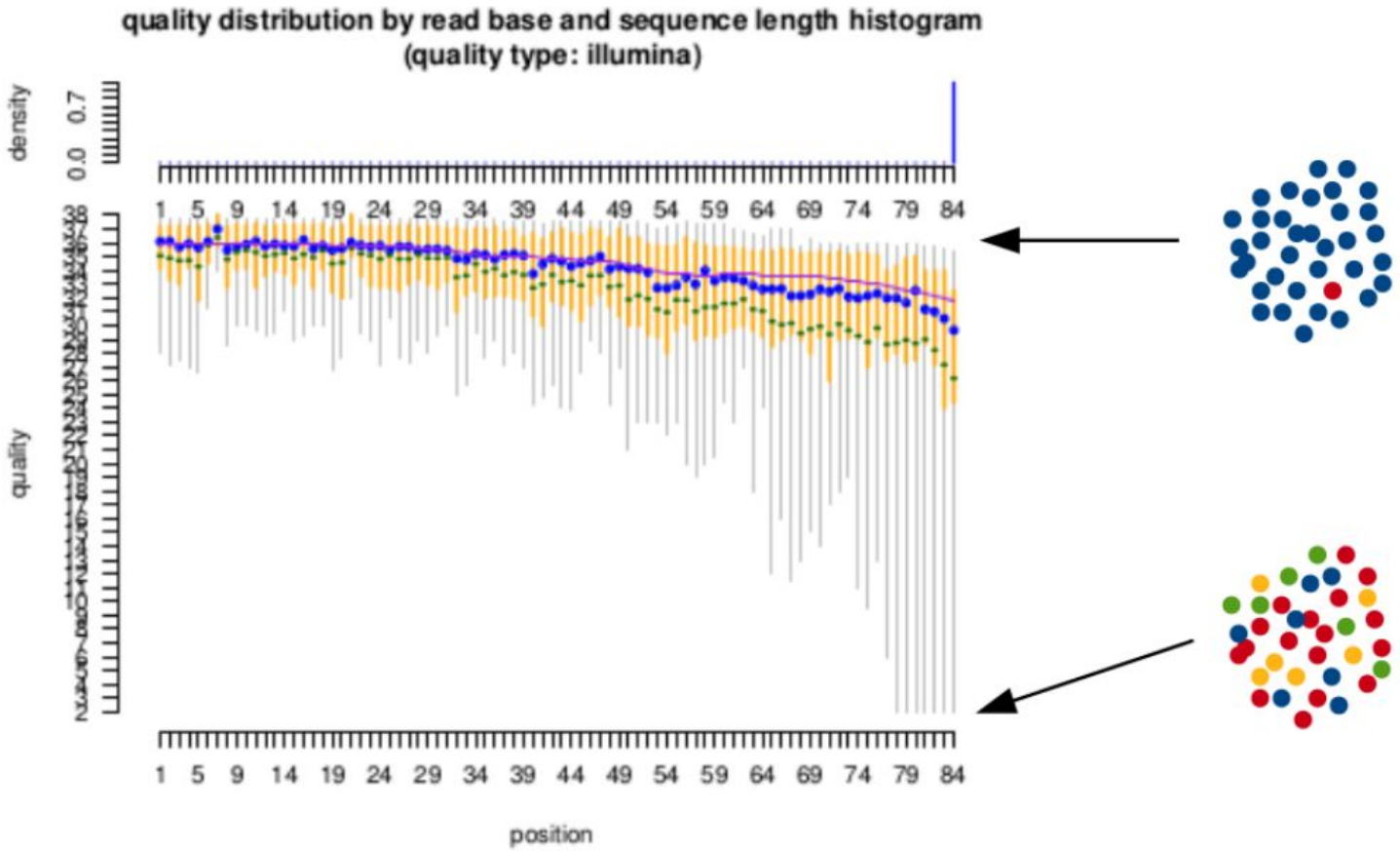
TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

- Asynchronizace molekul ve svazku (prephasing a postphasing)



http://bioinformatics.ucdavis.edu/docs/2015-september-workshop/_downloads/Monday_JF_QAI_lecture.pdf

- Asynchronizace molekul ve svazku (prephasing a postphasing)



http://bioinformatics.ucdavis.edu/docs/2015-september-workshop/_downloads/Monday_JF_QAI_lecture.pdf

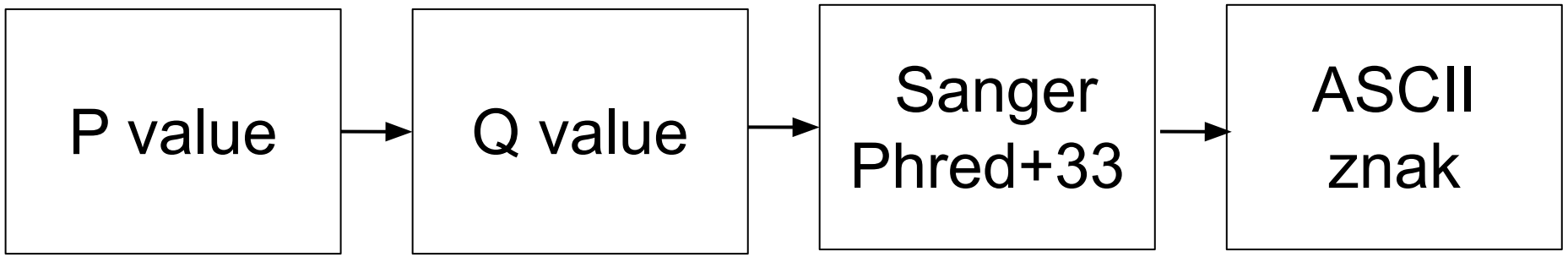
Q = –10 log(P-value)

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

FASTQ formát - kvalita bází

FG**I**EHCCD9DG=1E9?D>CF@HHG??B<GEBGHCG;;CDB8==C@@>>G**II**@@



0.0001	40	73	I

ASCII

Hex	Dec	Char		Hex	Dec	Char	Hex	Dec	Char	Hex	Dec	Char
0x00	0	NULL	null	0x20	32	Space	0x40	64	@	0x60	96	`
0x01	1	SOH	Start of heading	0x21	33	!	0x41	65	A	0x61	97	a
0x02	2	STX	Start of text	0x22	34	"	0x42	66	B	0x62	98	b
0x03	3	ETX	End of text	0x23	35	#	0x43	67	C	0x63	99	c
0x04	4	EOT	End of transmission	0x24	36	$	0x44	68	D	0x64	100	d
0x05	5	ENQ	Enquiry	0x25	37	%	0x45	69	E	0x65	101	e
0x06	6	ACK	Acknowledge	0x26	38	&	0x46	70	F	0x66	102	f
0x07	7	BELL	Bell	0x27	39	'	0x47	71	G	0x67	103	g
0x08	8	BS	Backspace	0x28	40	(	0x48	72	H	0x68	104	h
0x09	9	TAB	Horizontal tab	0x29	41	)	0x49	73	I	0x69	105	i
0x0A	10	LF	New line	0x2A	42	*	0x4A	74	J	0x6A	106	j
0x0B	11	VT	Vertical tab	0x2B	43	+	0x4B	75	K	0x6B	107	k
0x0C	12	FF	Form Feed	0x2C	44	,	0x4C	76	L	0x6C	108	l
0x0D	13	CR	Carriage return	0x2D	45	-	0x4D	77	M	0x6D	109	m
0x0E	14	SO	Shift out	0x2E	46	.	0x4E	78	N	0x6E	110	n
0x0F	15	SI	Shift in	0x2F	47	/	0x4F	79	O	0x6F	111	o
0x10	16	DLE	Data link escape	0x30	48	0	0x50	80	P	0x70	112	p
0x11	17	DC1	Device control 1	0x31	49	1	0x51	81	Q	0x71	113	q
0x12	18	DC2	Device control 2	0x32	50	2	0x52	82	R	0x72	114	r
0x13	19	DC3	Device control 3	0x33	51	3	0x53	83	S	0x73	115	s
0x14	20	DC4	Device control 4	0x34	52	4	0x54	84	T	0x74	116	t
0x15	21	NAK	Negative ack	0x35	53	5	0x55	85	U	0x75	117	u
0x16	22	SYN	Synchronous idle	0x36	54	6	0x56	86	V	0x76	118	v
0x17	23	ETB	End transmission block	0x37	55	7	0x57	87	W	0x77	119	w
0x18	24	CAN	Cancel	0x38	56	8	0x58	88	X	0x78	120	x
0x19	25	EM	End of medium	0x39	57	9	0x59	89	Y	0x79	121	y
0x1A	26	SUB	Substitute	0x3A	58	:	0x5A	90	Z	0x7A	122	z
0x1B	27	FSC	Escape	0x3B	59	;	0x5B	91	[	0x7B	123	{
0x1C	28	FS	File separator	0x3C	60	<	0x5C	92	\	0x7C	124	|
0x1D	29	GS	Group separator	0x3D	61	=	0x5D	93	]	0x7D	125	}
0x1E	30	RS	Record separator	0x3E	62	>	0x5E	94	^	0x7E	126	~
0x1F	31	US	Unit separator	0x3F	63	?	0x5F	95	_	0x7F	127	DEL

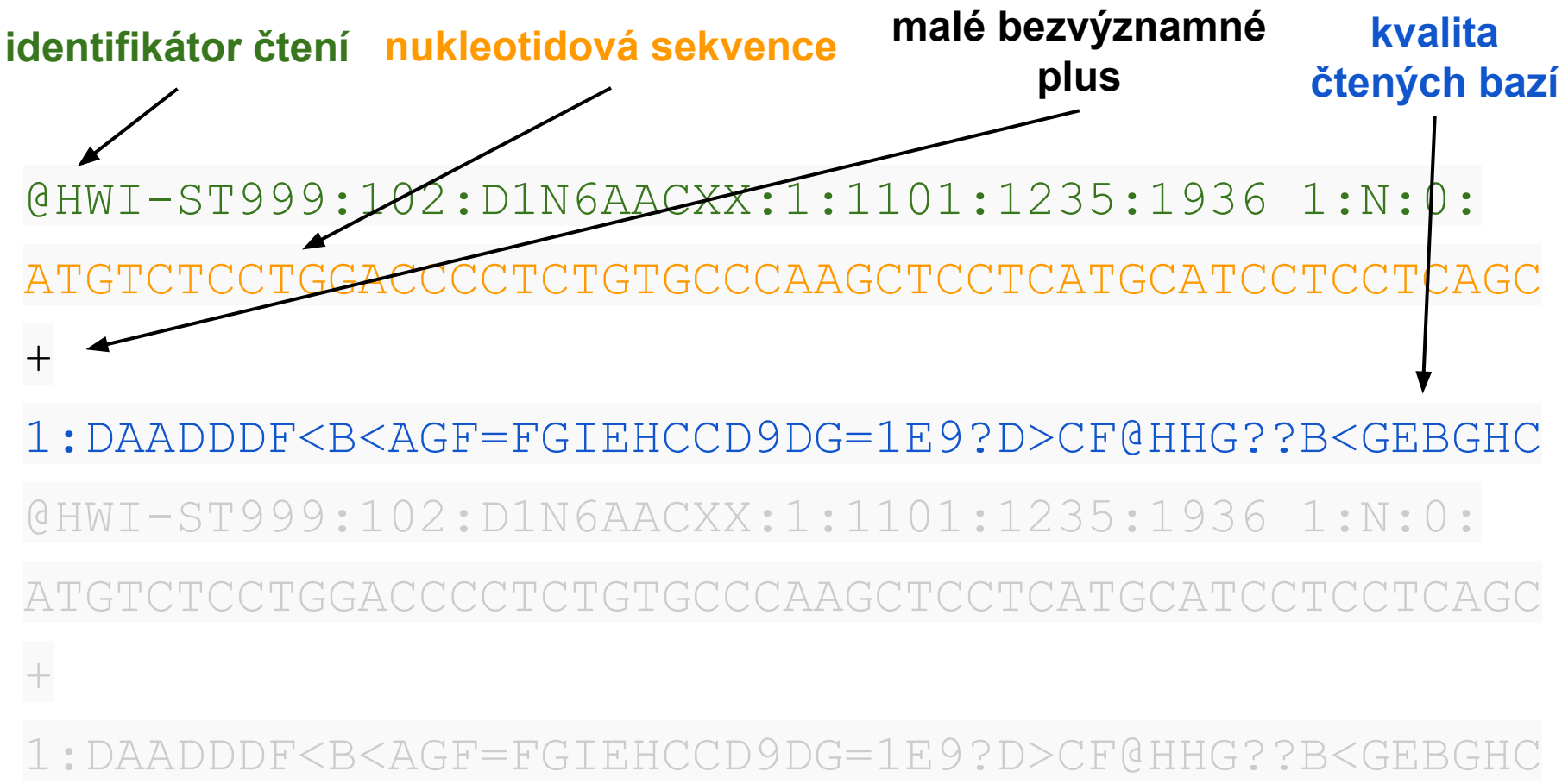
Různé formáty kvality bází

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |       |                              |                     |
33                        59   64      73                             104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

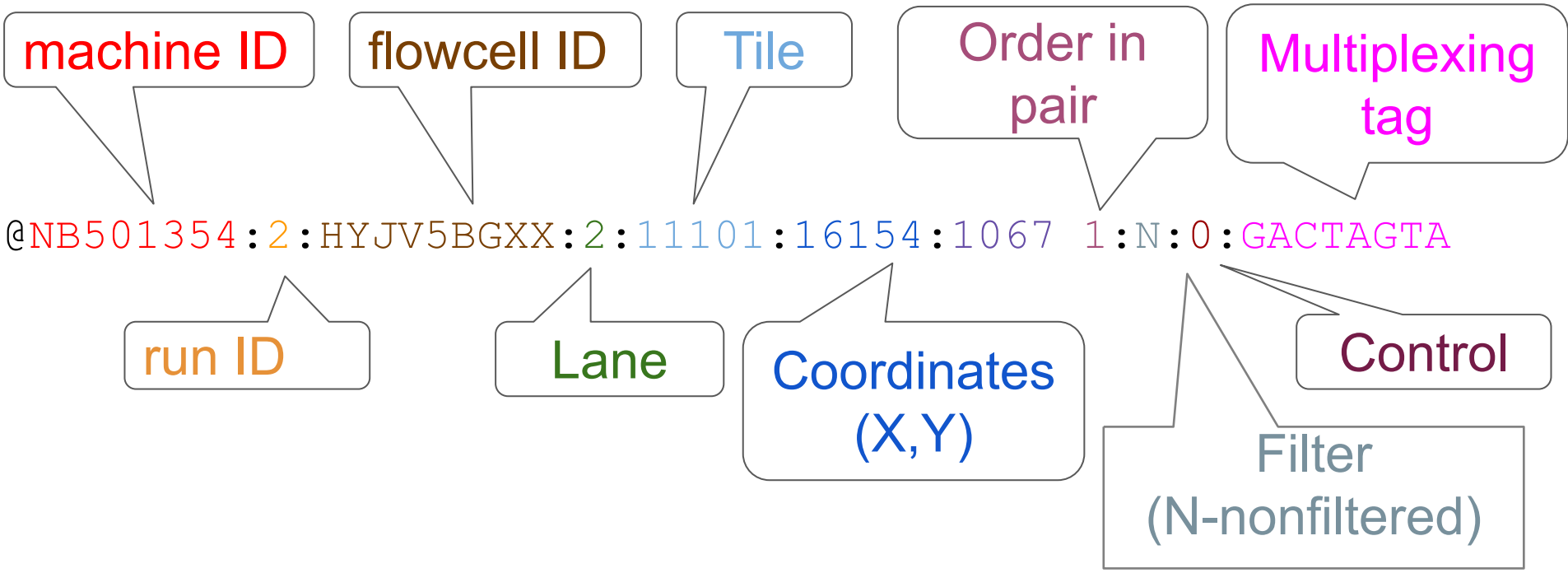
https://en.wikipedia.org/wiki/FASTQ_format

FASTQ file

Formát pro zápis sekvenčních dat - primární výstup sekvenace

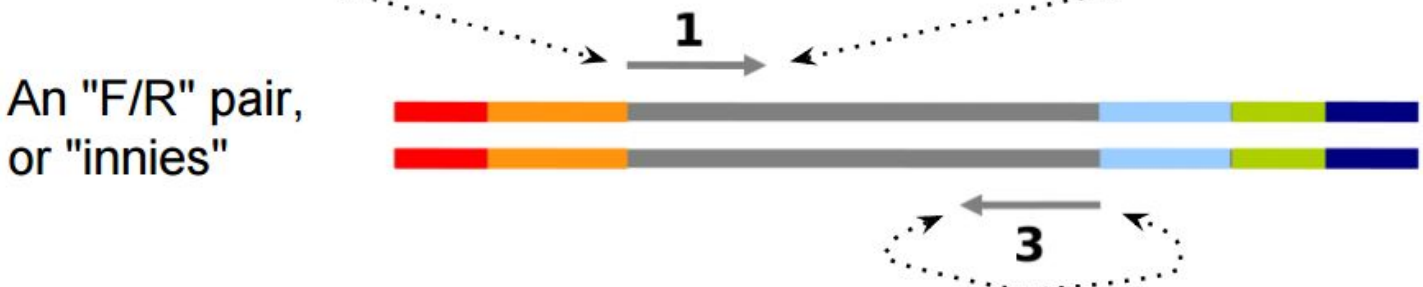


FASTQ format - read identifier



Orientace čtení ve FASTQ

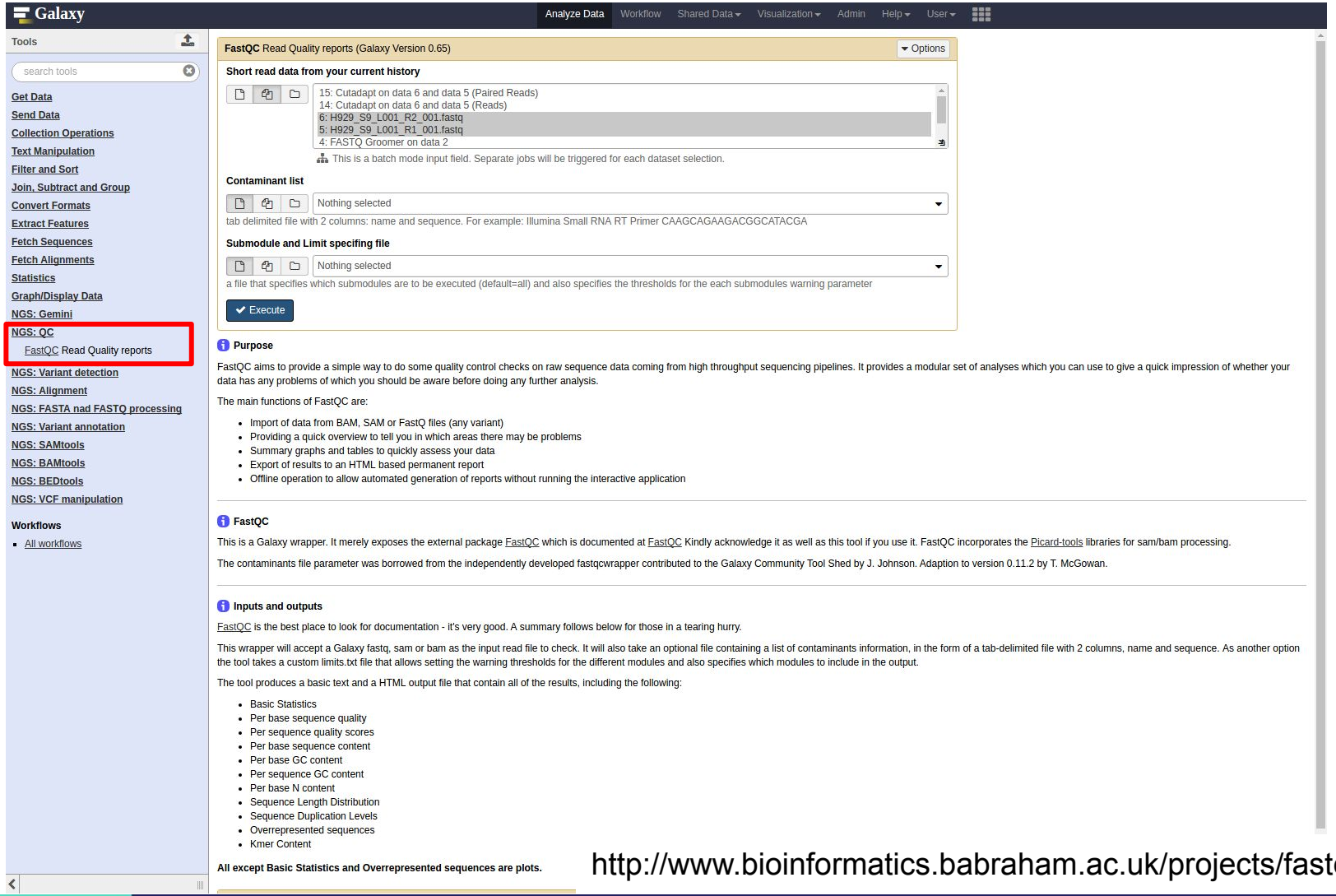
```
@HWI-700593F:508:H2FGCBCXX:1:1101:5750:2167 1:N:0:AGTCAA
AAATGAGATGAAATTCGTTCAAGCCAAAGGAGAGAGCCTTTCTTCAACATTTCTTCCATTTTCTCTCAAATCATCCAT
+
BDADBECHHFDECHIHIFGHFFHIIIIHHFHIHHGHHHIIHHGGHEGGGI1FGGHHHEHH<FCGHEEEFCHICHHCCF
```




```
@HWI-700593F:508:H2FGCBCXX:1:1101:5750:2167 2:N:0:AGTCAA
CATCAAAACTAAAGTTTTTAATCCTTACCTTAAACCCTTCAAATTGAAGATCCAAGGCTCCTATCTTGTTCTCCTTGA
+
DADD@HHHFHHHHH1CG?GHFFHHIIIIIIIIIHHHHHHHHHHHIIEHHHHGGHHIGIIHIGIIIIHIIHHHI?H11C
```

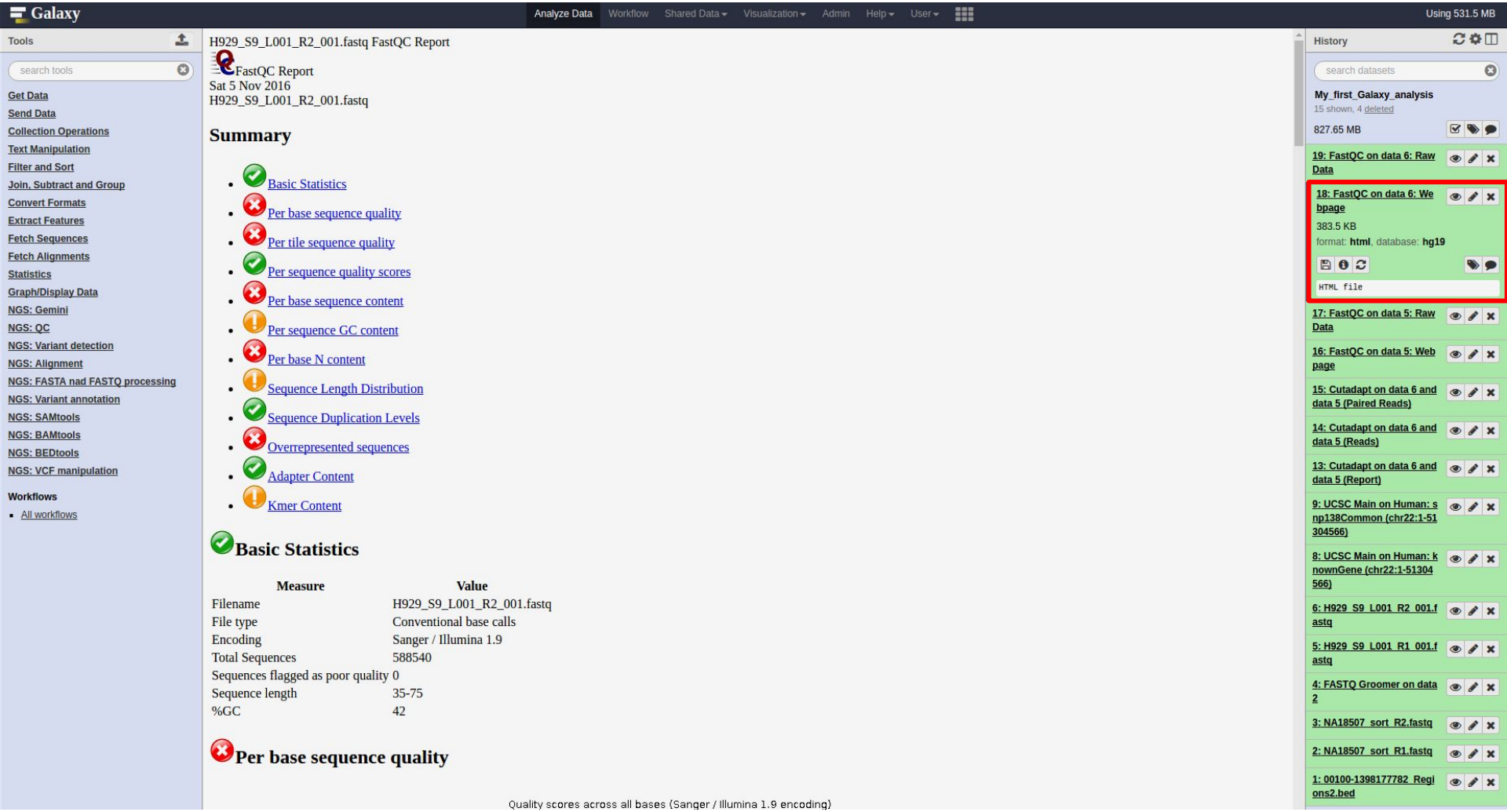
http://bioinformatics.ucdavis.edu/docs/2015-september-workshop/_downloads/Monday_JF_QAI_lecture.pdf

NGS:QC -> FastQC



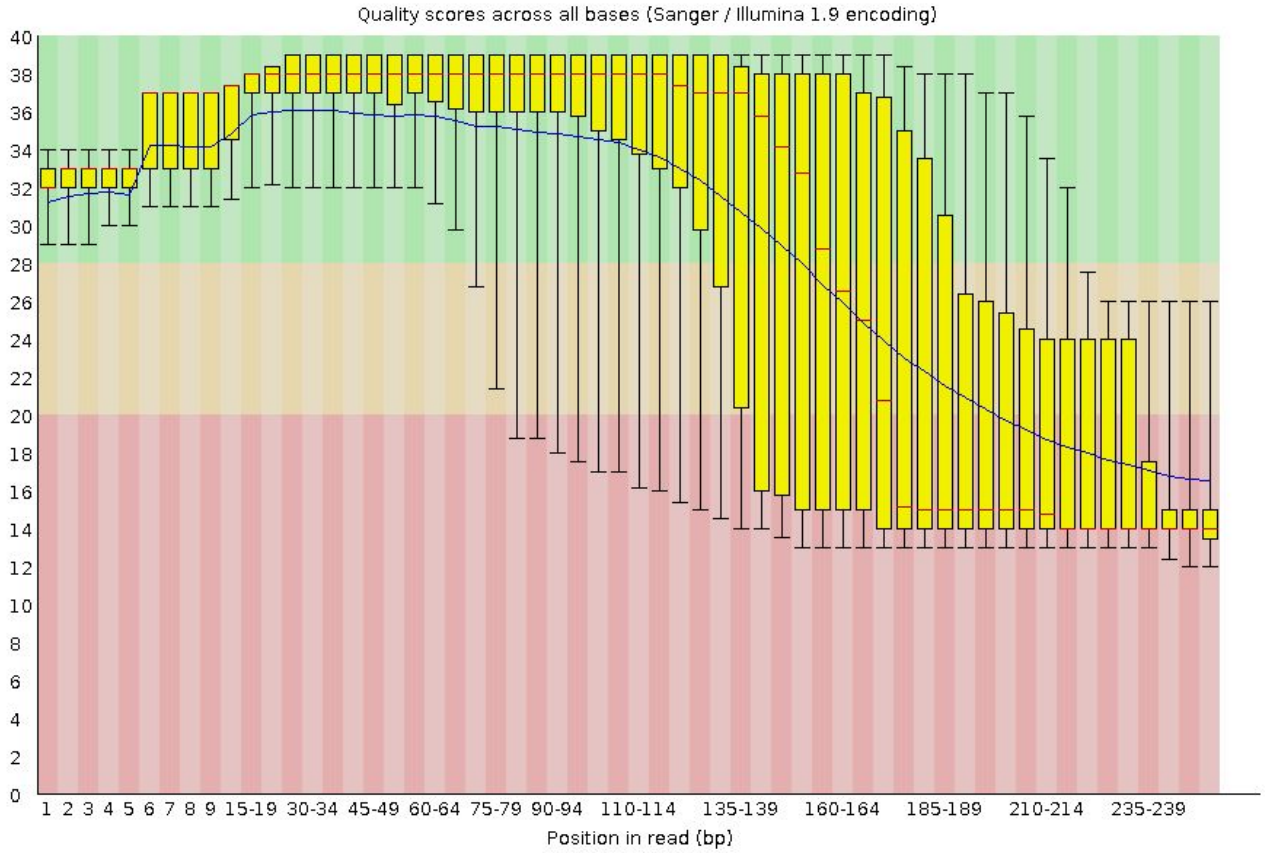
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

NGS:QC -> FastQC

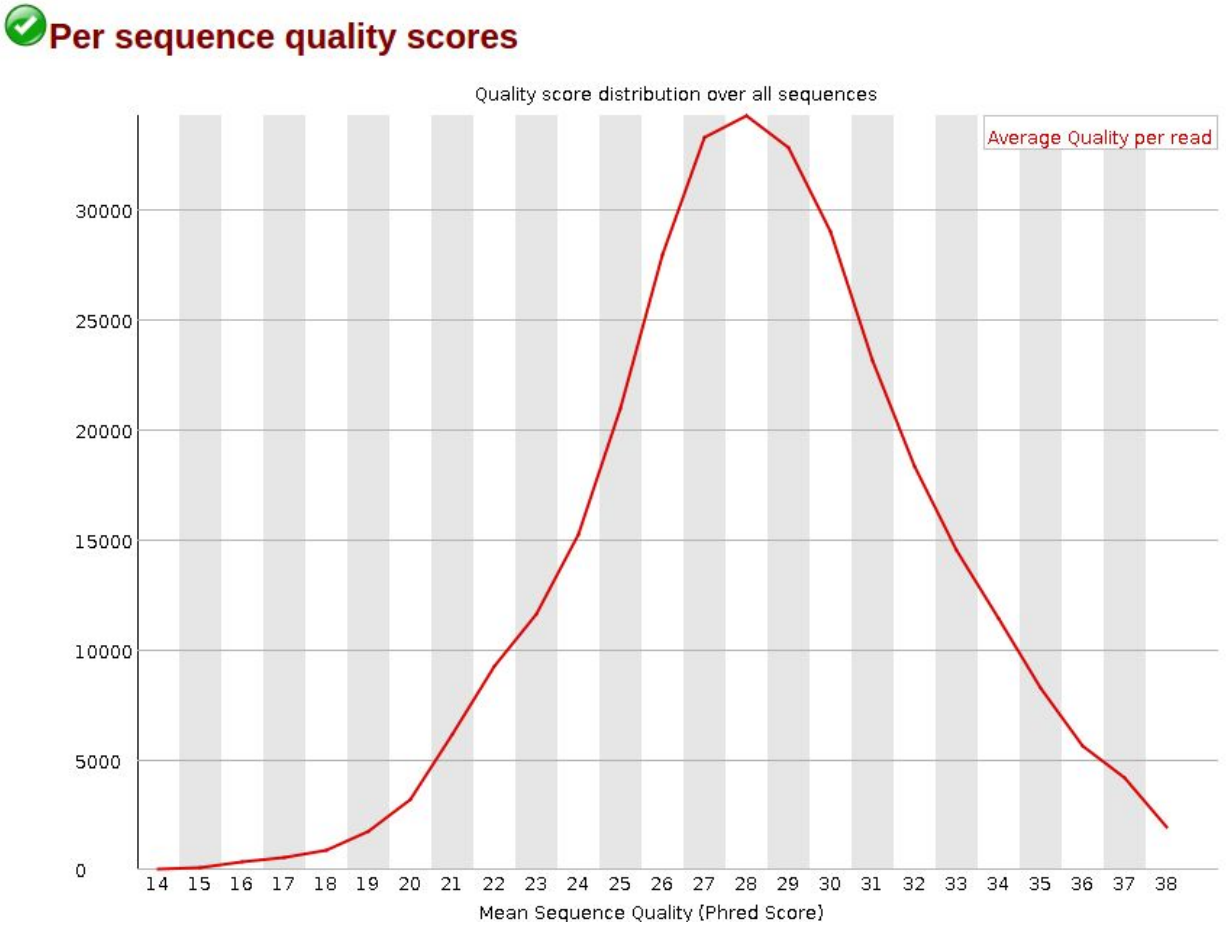


Kvalita bází

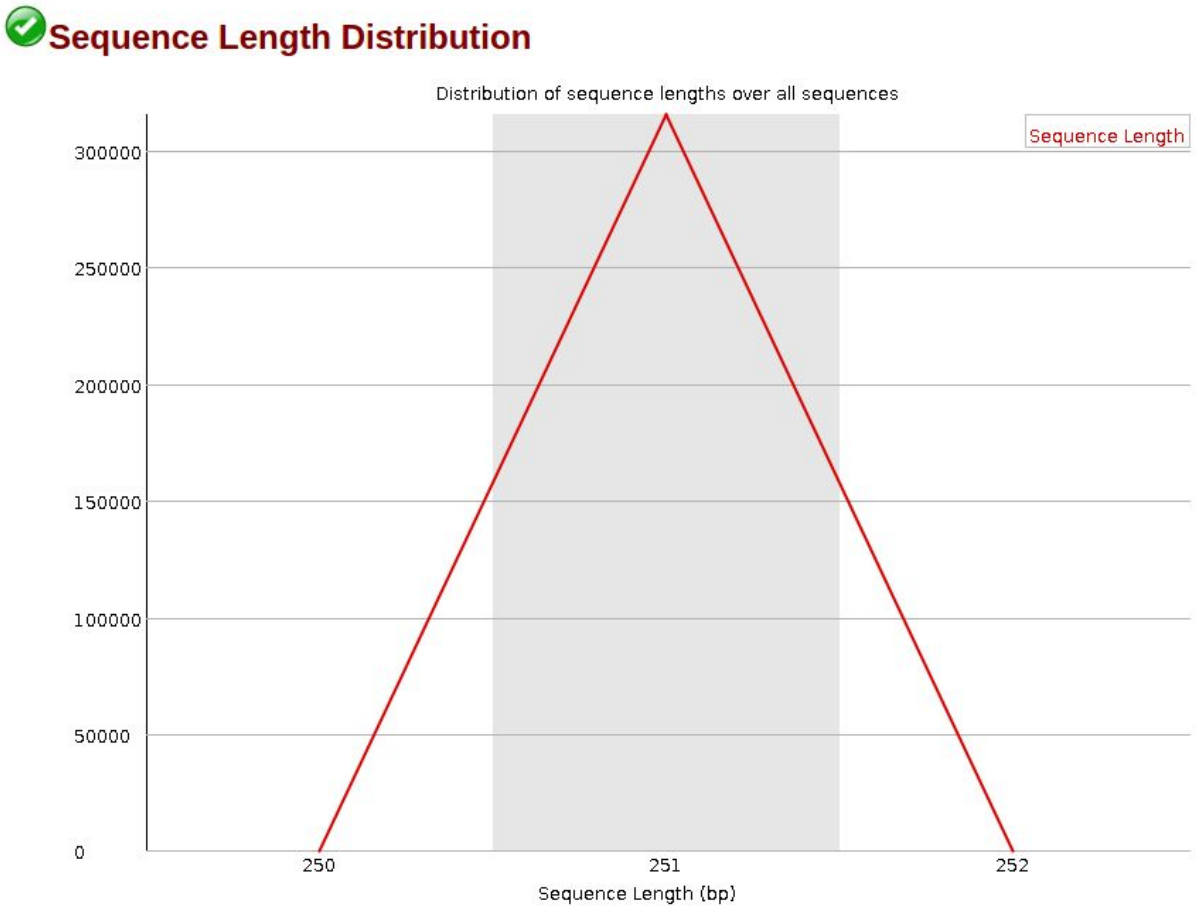
Per base sequence quality



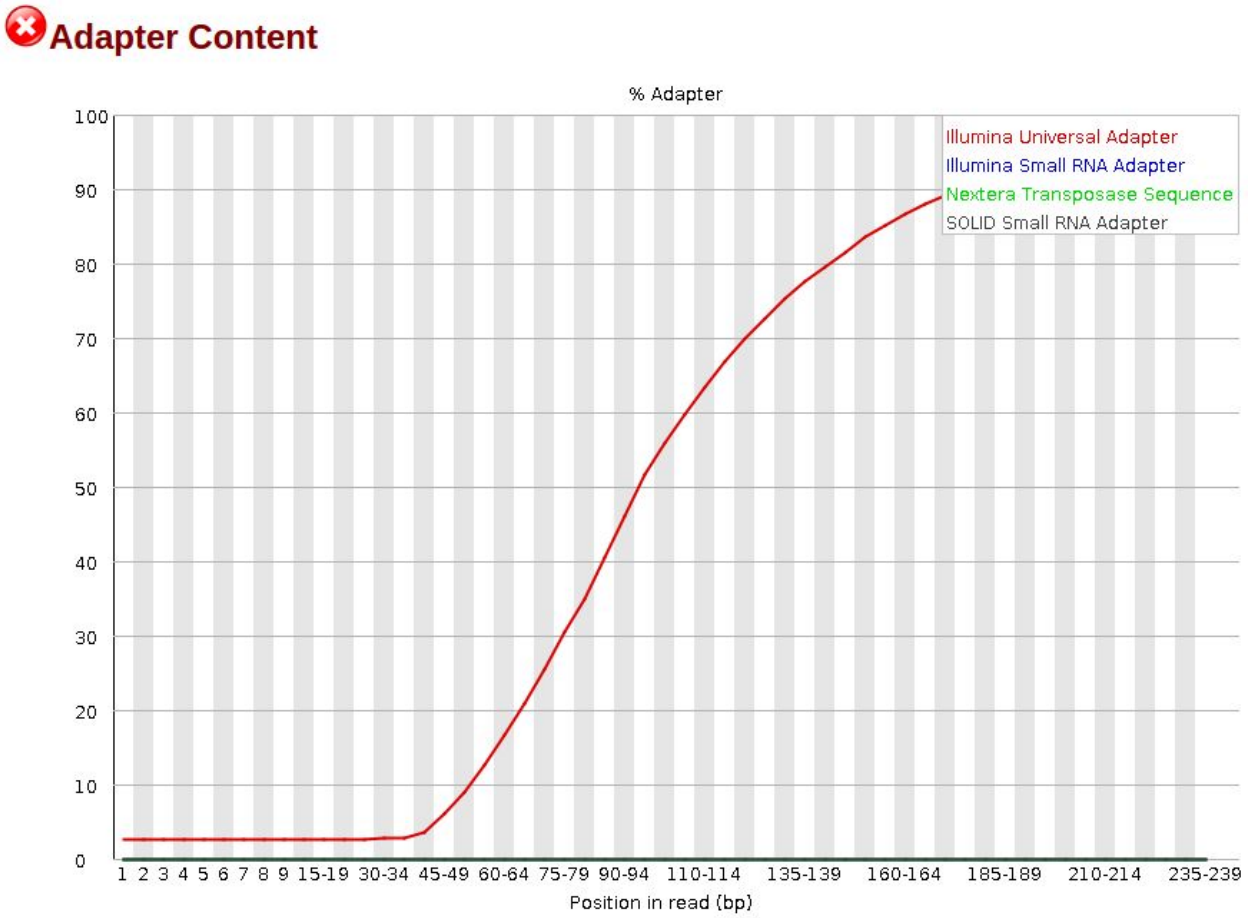
Kvalita čtení



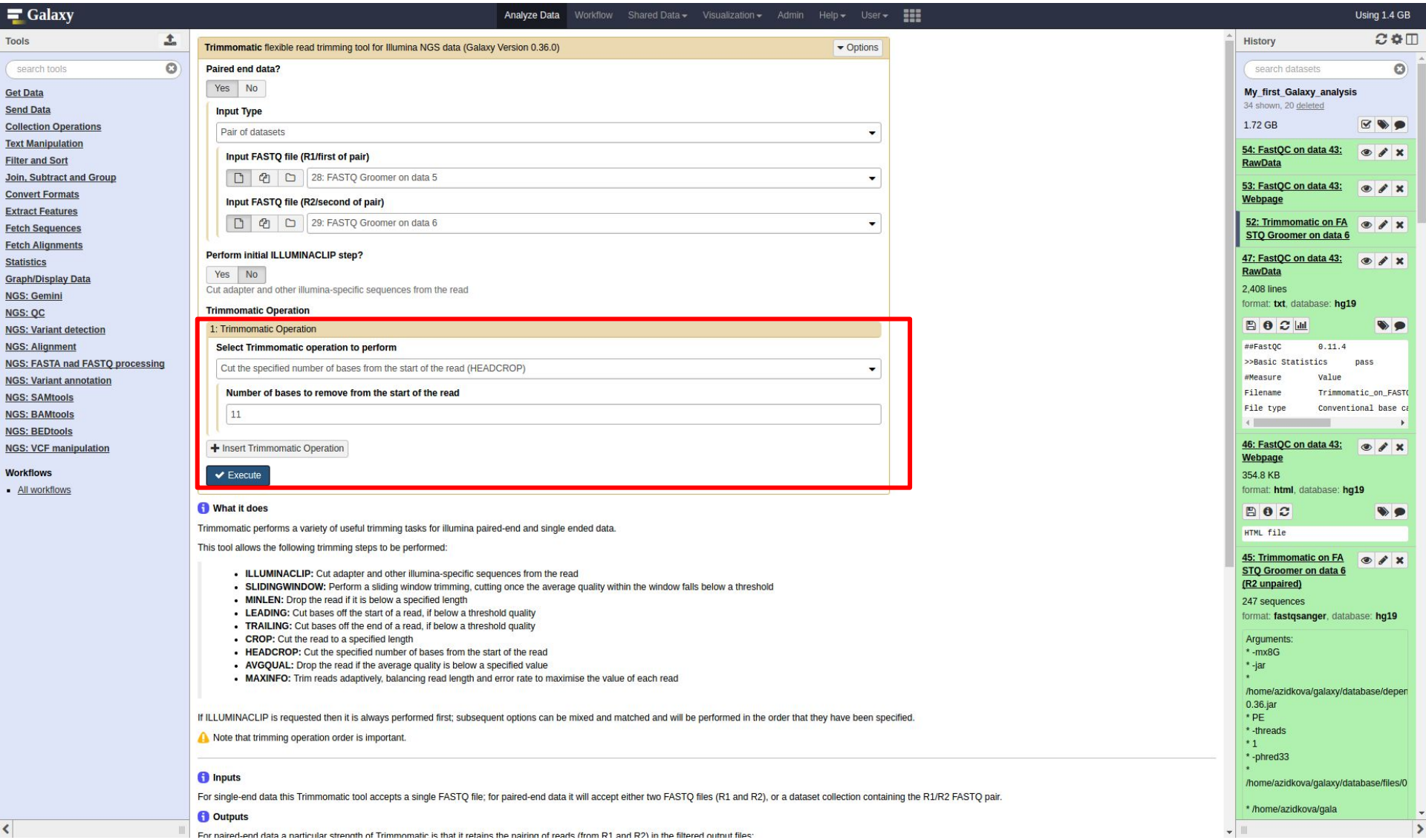
Délka čtení


Sequence Length Distribution
Distribution of sequence lengths over all sequences
Sequence Length
300000
250000
200000
150000
100000
50000
0
250
251
252
Sequence Length (bp)

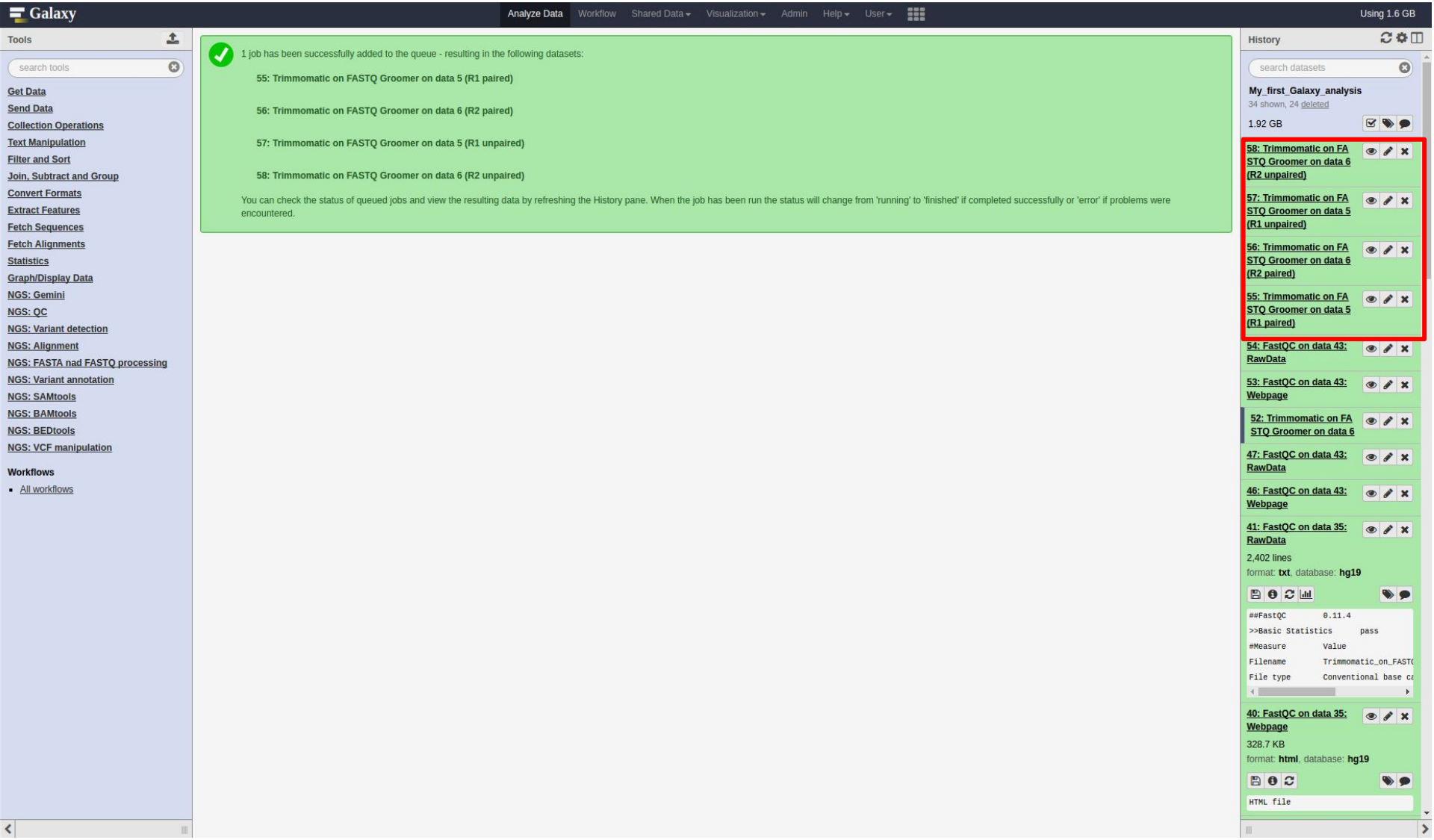
Kontaminace adaptéry



Ořezání dat - Trimmomatic



Ořezání dat - Trimmomatic

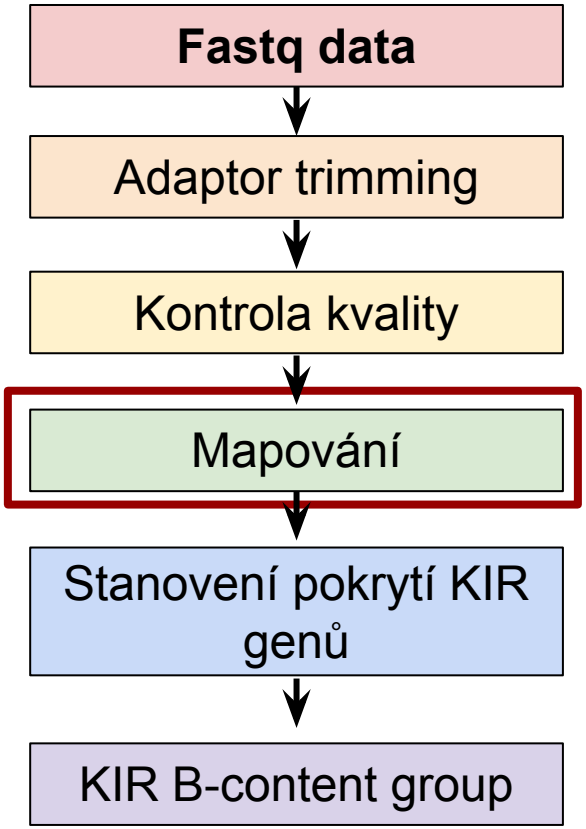


Cvičení 1 - kontrola kvality a trimming

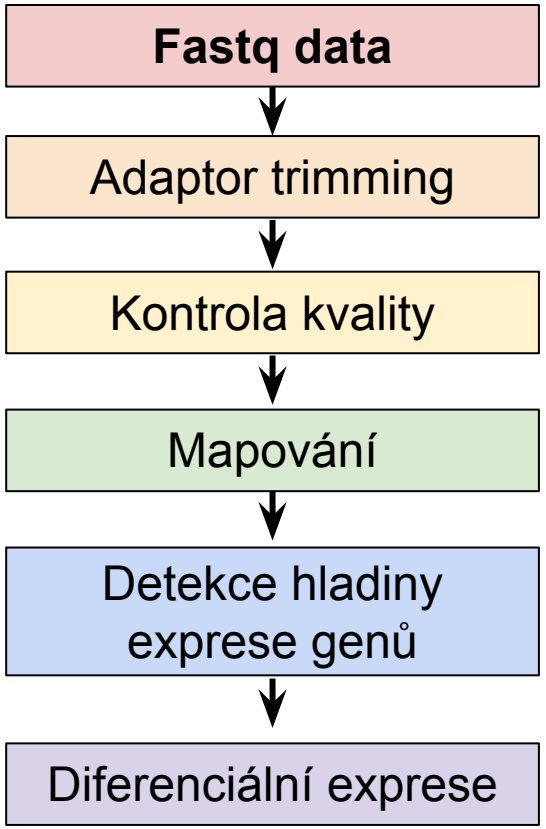
1. V Galaxy importujte historii "**KIR-*QC***" (https://usegalaxy.org/u/tomas_hron/h/kir-qc).
2. Spusťte kontrolu kvality FastQ souborů pomocí nástroje "***FastQC***". **Byl u dat proveden "adaptor trimming"?**
3. Filtrujte data pomocí "***Trimmomatic***" (IlluminaClip, TruSeq3 PE adaptory;TRAILING, kvalita>20).
4. Proveďte kontrolu kvality na ořezaných datech dle bodu 3. **Jaká je minimální délka čtení? Kolik procent párových čtení nám zbylo z původního počtu?**

***https://usegalaxy.org/u/tomas_hron/h/kir-qc-finished**

Genotypizace KIRs



Diferenciální exprese genů u D. melanogaster



The first printout of the human genome to be presented as a series of books, displayed at the Wellcome Collection, London

Genome reference sequence

Release name	Date of release	Equivalent UCSC version
GRCh38	Dec 2013	hg38
GRCh37	Feb 2009	hg19
NCBI Build 36.1	Mar 2006	hg18
NCBI Build 35	May 2004	hg17
NCBI Build 34	Jul 2003	hg16

Genome reference sequence

GRCh37.p13:

Total bases:

3.23 Billion
2.99 Billion (without N)

N50:

46 Million

Number of alternative loci:

9

Non-nuclear genome:

No

GRCh38.p2:

Total bases:

3.21 Billion
3.05 Billion (without N)

N50:

67 Million

Number of alternative loci :

261

Non-nuclear genome:

Yes

Reference transcriptome sequences

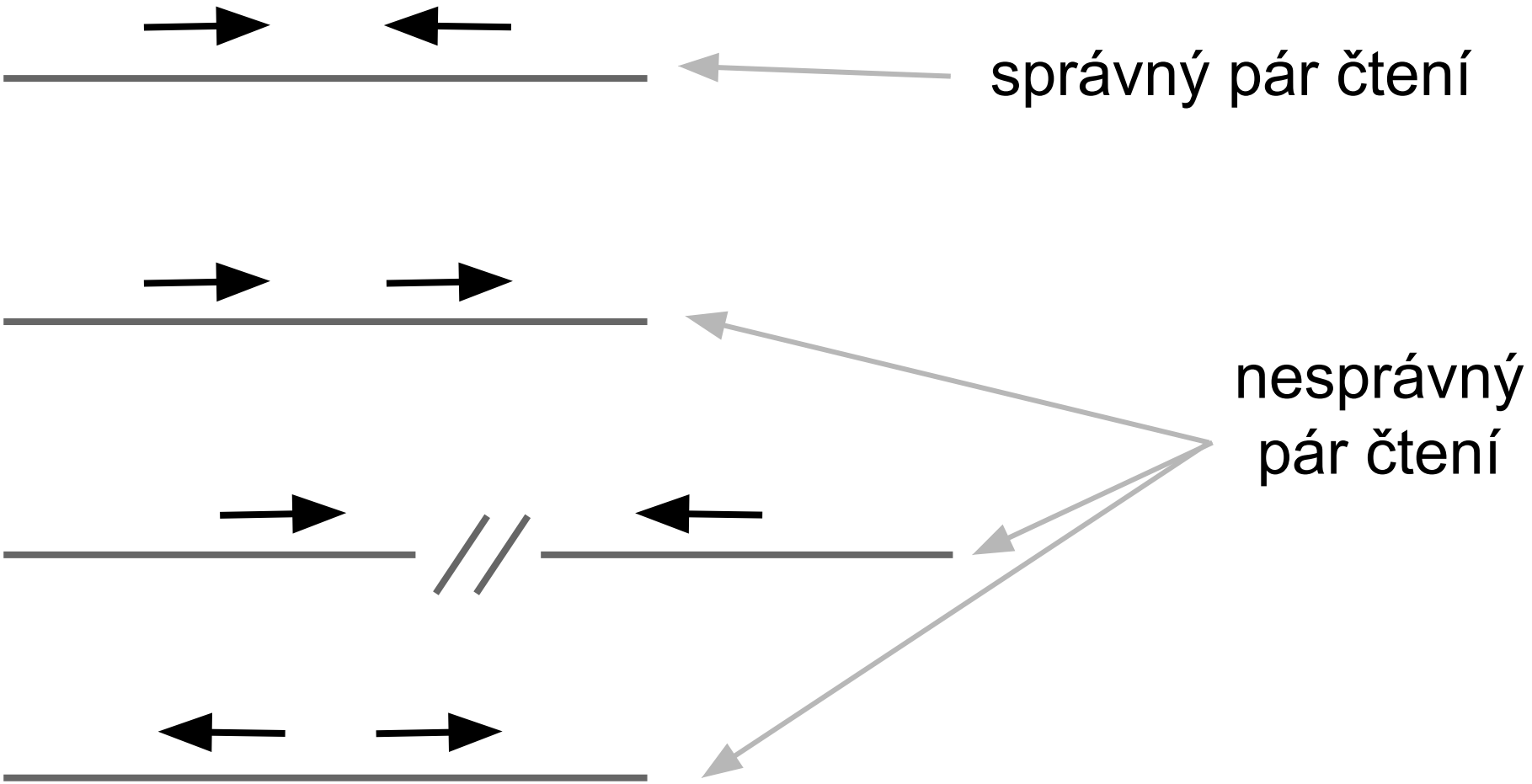
RefSeq

The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form a foundation for medical, functional, and diversity studies. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially RefSeqGene records), expression studies, and comparative analyses.

GENCODE

For human and mouse, this combined Ensembl/HAVANA gene set is the default gene set from the GENCODE project. For both human and mouse, we also guarantee that all transcripts from the Consensus Coding Sequence (CCDS) set are present in the GENCODE gene set.
The complete GENCODE gene set for human and mouse can be viewed on our website in the GENCODE Comprehensive track. A subset of these transcript models are tagged as being in the GENCODE Basic set, which is also available on our website.

Párová čtení (paired-end reads)



Edit distance

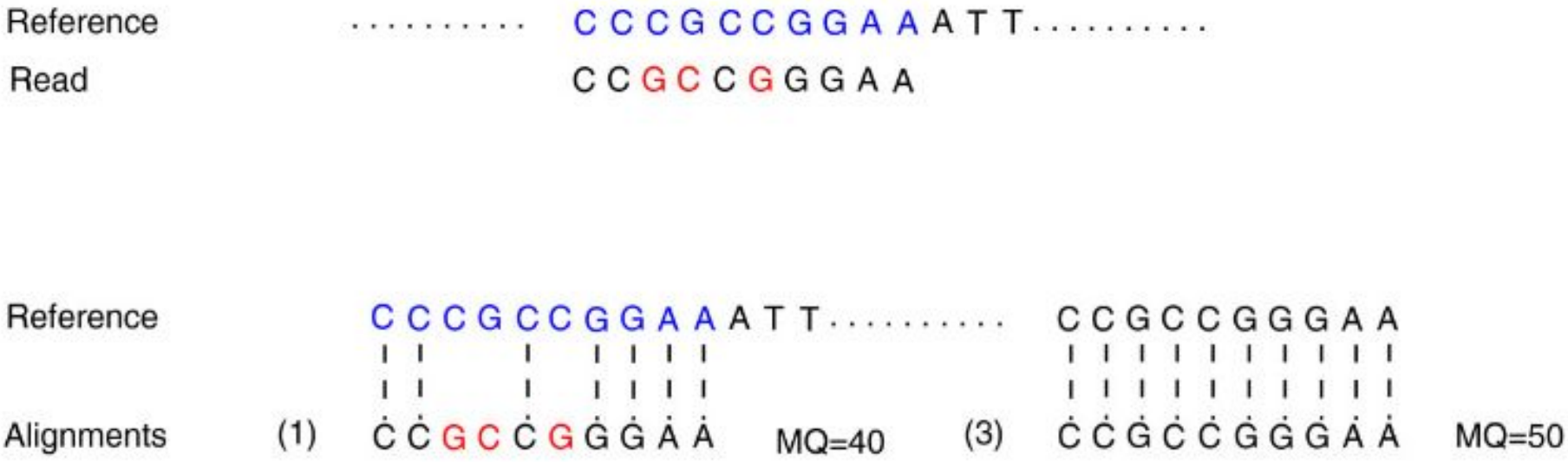
```
AACGT-AGCC Reference
A--GTCATCC Čtení

Edit distance = 4
```

Mapping quality

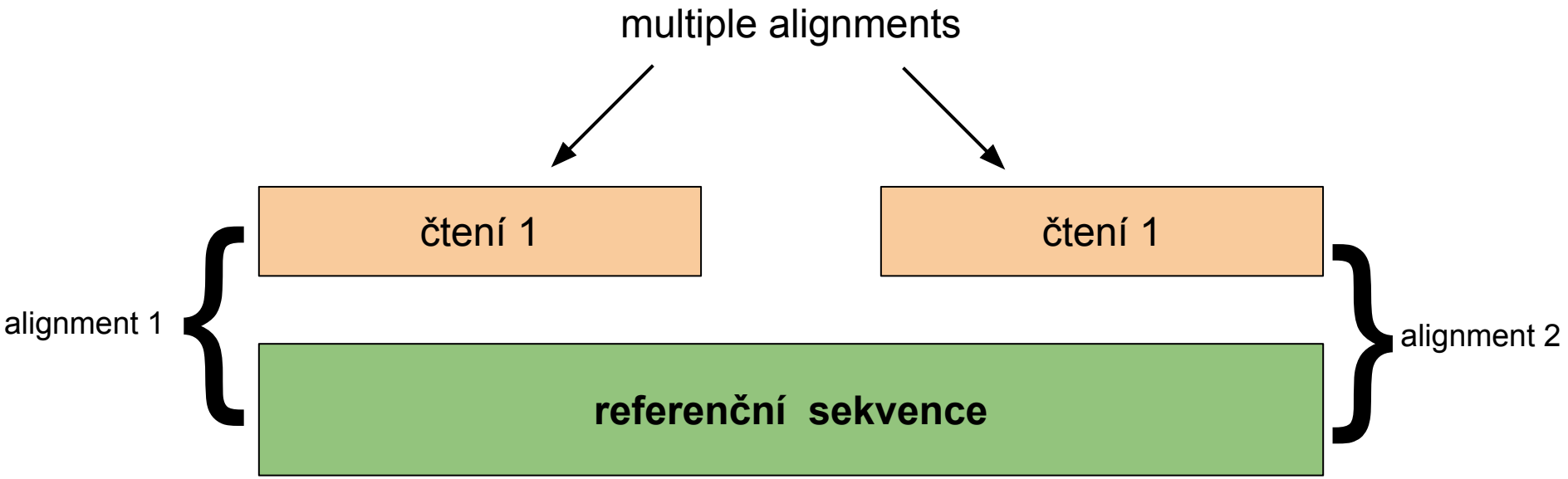
The calculation of mapping qualities considers all the factors below:

- The repeat structure of the reference. Reads falling in repetitive regions usually get very low mapping quality.
- The base quality of the read. Low quality means the observed read sequence is possibly wrong, and wrong sequence may lead to a wrong alignment.
- Paired end or not. Reads mapped in pairs are more likely to be correct.

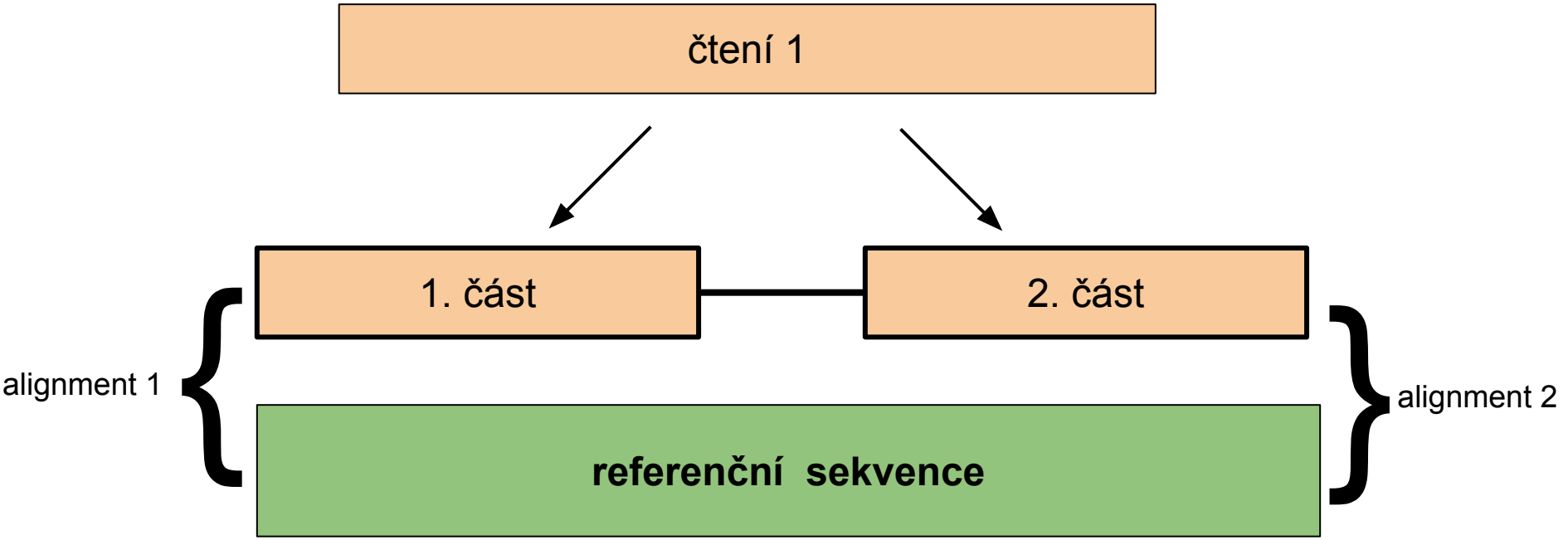
Mapping quality



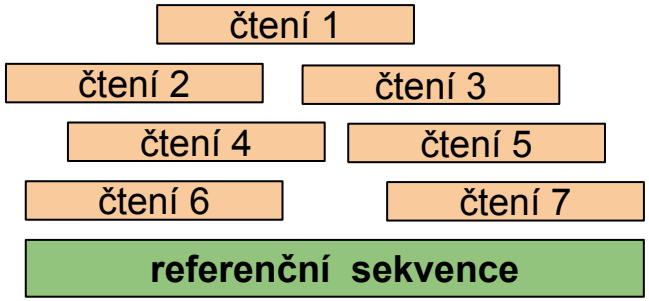
secondary alignment



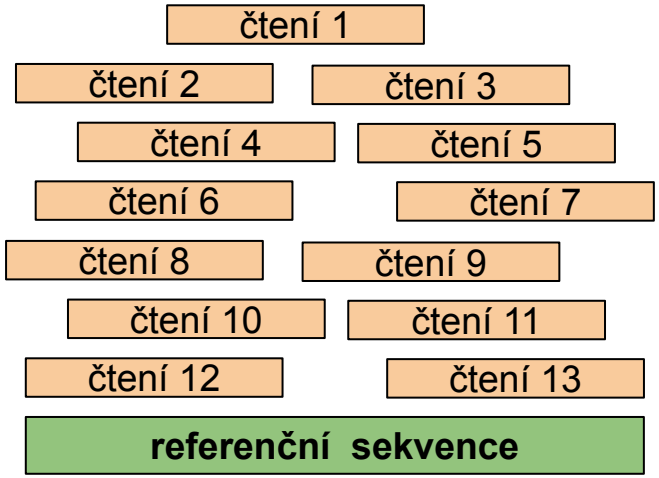
split alignment



Hloubka sekvenačního pokrytí

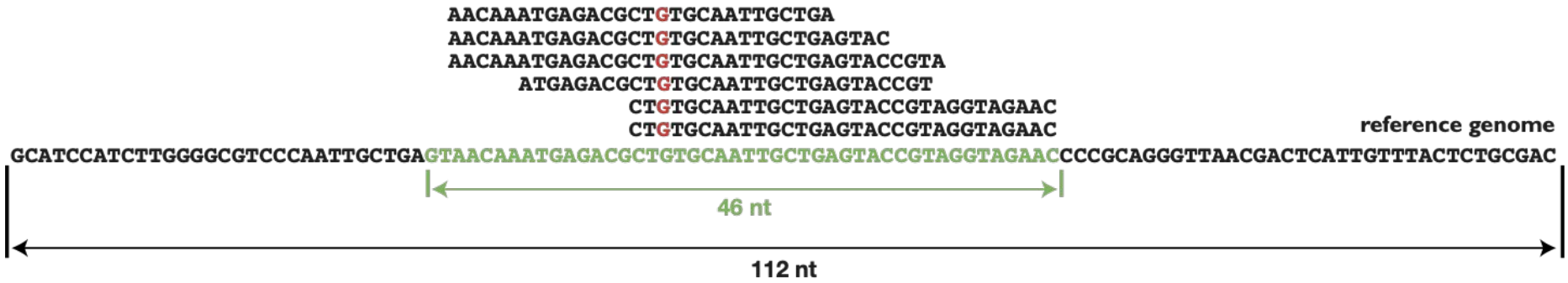


nízké pokrytí
(low coverage)



vysoké pokrytí
(high coverage)

Hloubka sekvenačního pokrytí



Reference point	Calculation	Example (see Fig.2)
Whole genome	(# of sequenced bases)* / (genome size)	188 / 112 = 1,68 fold
One locus	(# of bases mapping to the locus) / (size of locus)	188 / 46 = 4,09 fold
One position	(# of reads overlapping with one position)	6 fold

SAM/BAM (Sequence/Binary Alignment Map)

- Hlavička (řádky začínají s @)
  - HD - header
  - SQ - reference sequence dictionary
  - RG - read group
  - PG - program
- Tabulka (1 read alignment na každý řádek)
  - QNAME - název čtení
  - FLAG - informace ohledně mapování
  - RNAME - název referenční sekvence
  - POS - pozice, na kterou bylo namapované čtení (nejvíce vlevo)
  - MAPQ - kvalita namapování čtení na referenční sekvenci (phred)
  - CIGAR - zapis alignmentu
  - RNEXT - název párového čtení (pokud stejny tak “=”, pokud neni - “*”)
  - PNEXT - pozice párového čtení (pokud není tak “0”)
  - TLEN - délka templátu / sekv. fragmentu (pokud není tak “0”)
  - SEQ - sekvence čtení
  - QUAL - kvalita čtení
  - TAGS - nepovinná doplňující data
- Více informací: https://samtools.github.io/hts-specs/SAMv1.pdf

SAM - příklad

hlavička

```
@HD VN:1.0  SO:coordinate
@SQ SN:1  LN:249250621  AS:NCBI37
UR:file:/data/GATK/human_g1k_v37.fasta
@RG ID:UM0098:1  PL:ILLUMINA  PU:HWUSI-L001  LB:80 ...
@PG ID:bwa  VN:0.5.4
```

alignment 1

```
1:497:R:-272+13M17D24M   113 1   497 37  37M 15  100338662   0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>> XT:A:U  NM:i:0  SM:i:37 AM:i:0
X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:37
```

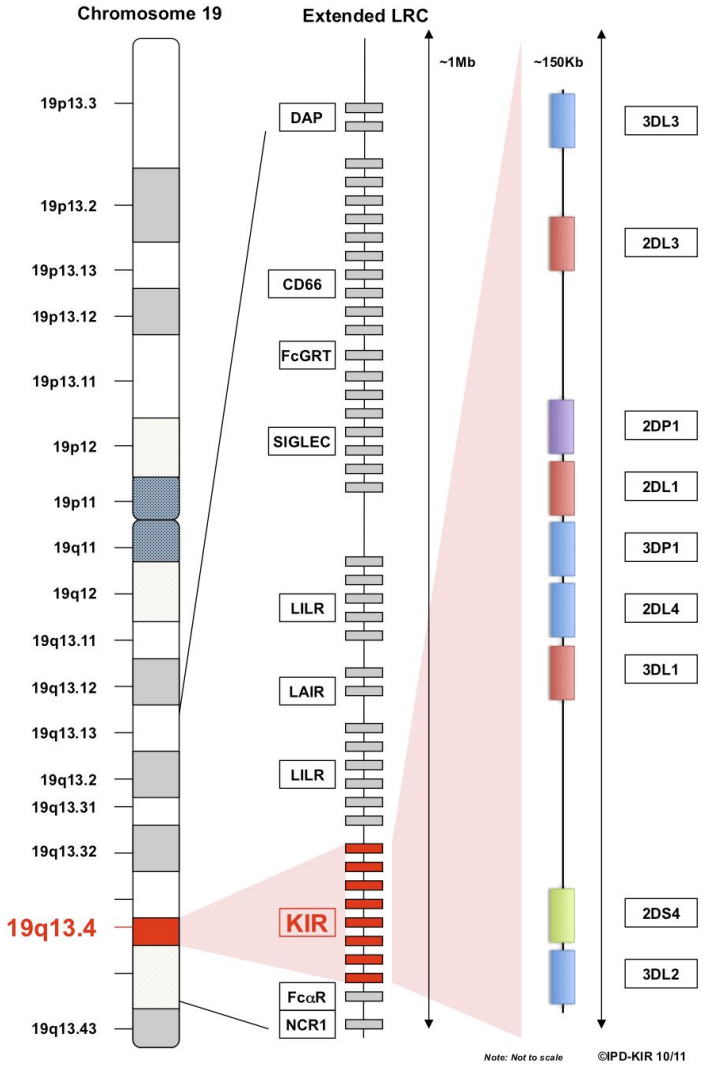
alignment 2

```
19:20389:F:275+18M2D19M  99  1   17644   0   37M =   17919   314
TATGACTGCTAATAATACCTACACATGTTAGAACCAT
>>>>>>>>>>>>>>>>>>>><>>><<>>4:>>:<9 RG:Z:UM0098:1   XT:A:R  NM:i:0
SM:i:0  AM:i:0  X0:i:4  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:37
```

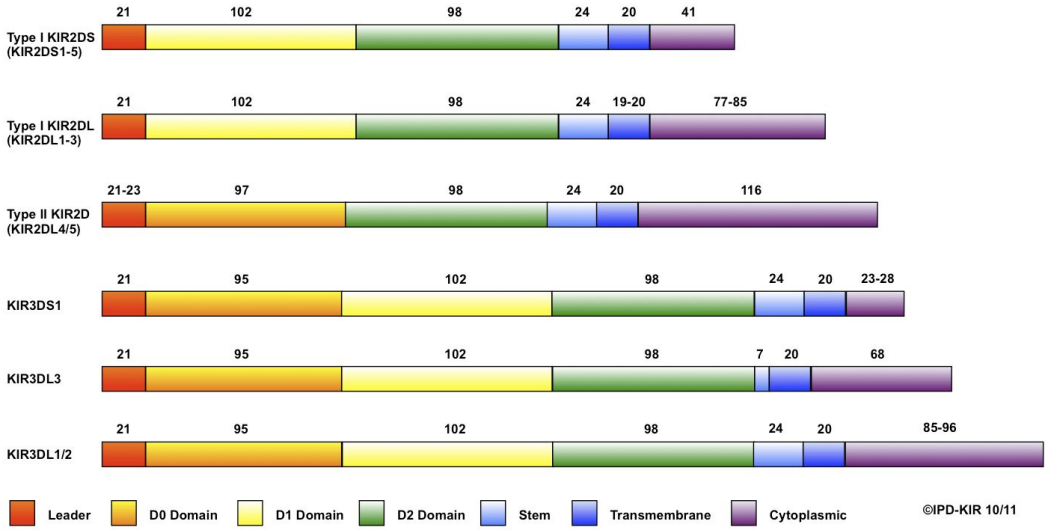
Úskalí mapování NGS dat

- Chybovost sekvenování
  - Illumina ~0.1%
- Variabilita
  - individual genome differs from the reference human genome at 4.1 million to 5.0 million sites affecting 20 million bases (1000 Genomes Project)
- Repetitivní oblasti
  - ~50% genomu

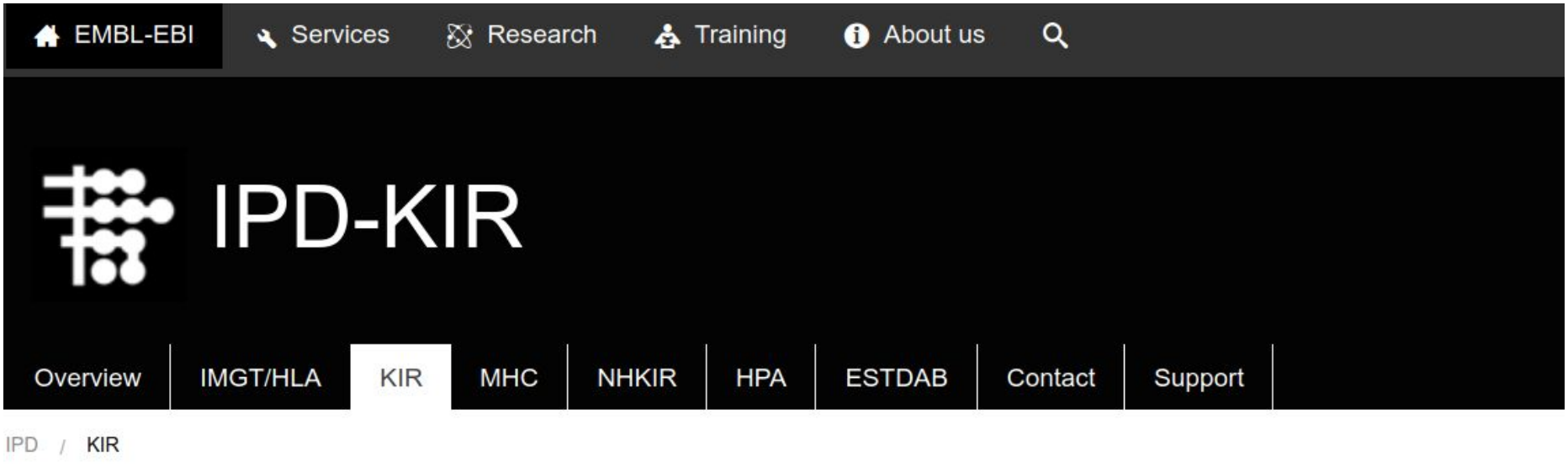
Jakou referenci použít pro KIRs?



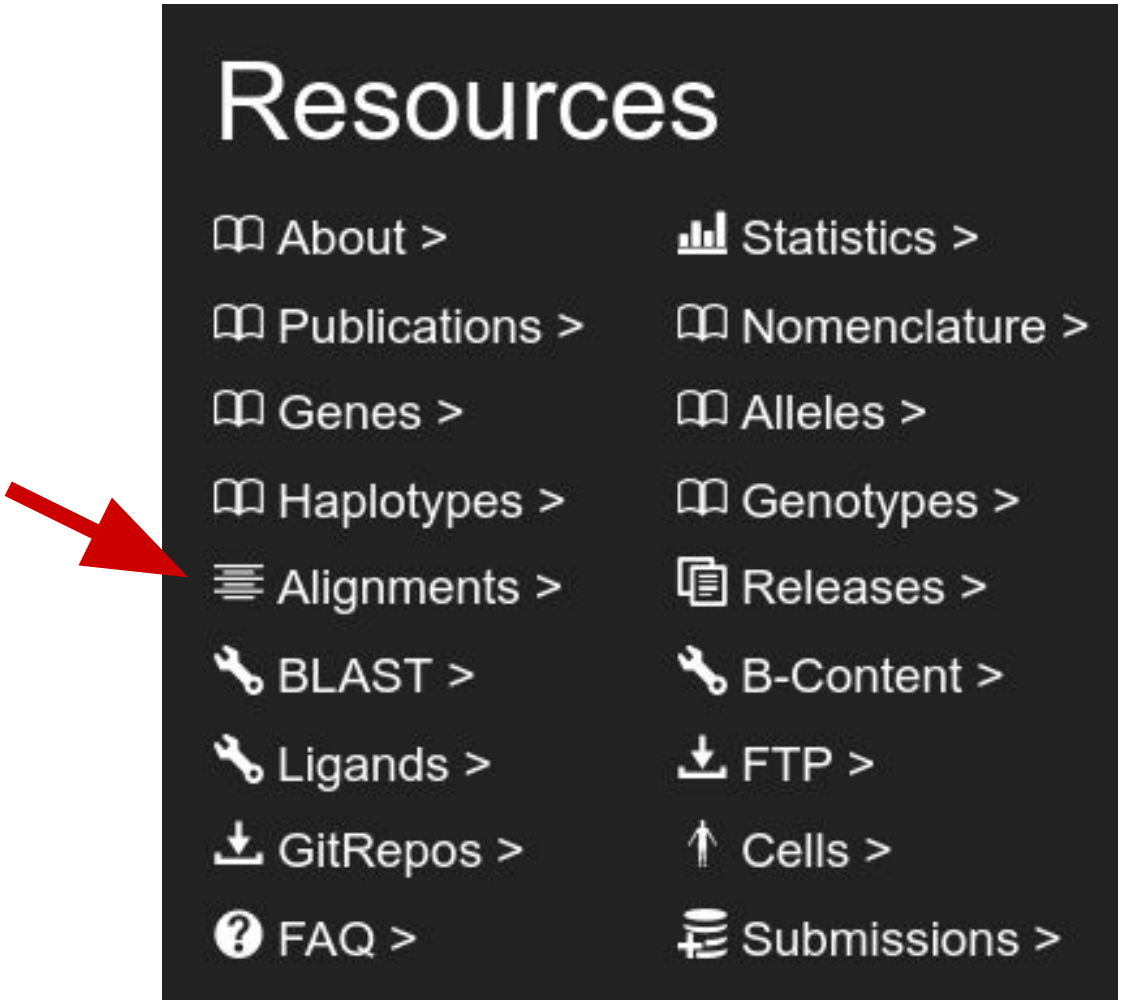
The KIR gene family currently consists of 15 gene loci (KIR2DL1, KIR2DL2/L3, KIR2DL4, KIR2DL5A, KIR2DL5B, KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS4, KIR2DS5, KIR3DL1/S1, KIR3DL2, KIR3DL3 and two pseudogenes, KIR2DP1 and KIR3DP1) encoded within a 100-200 Kb region of the Leukocyte Receptor Complex (LRC) located on chromosome 19 (19q13.4) (1).



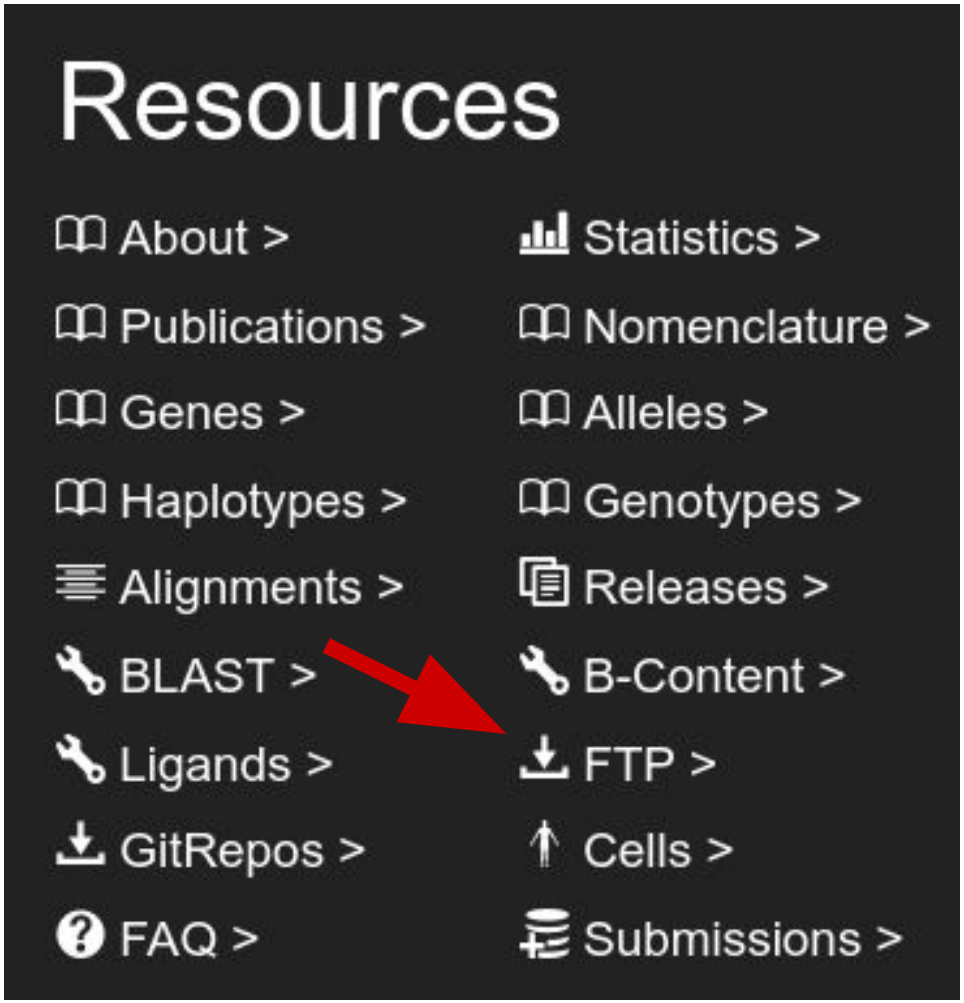
Jakou referenci použít pro KIRs?



Jakou referenci použít pro KIRs?

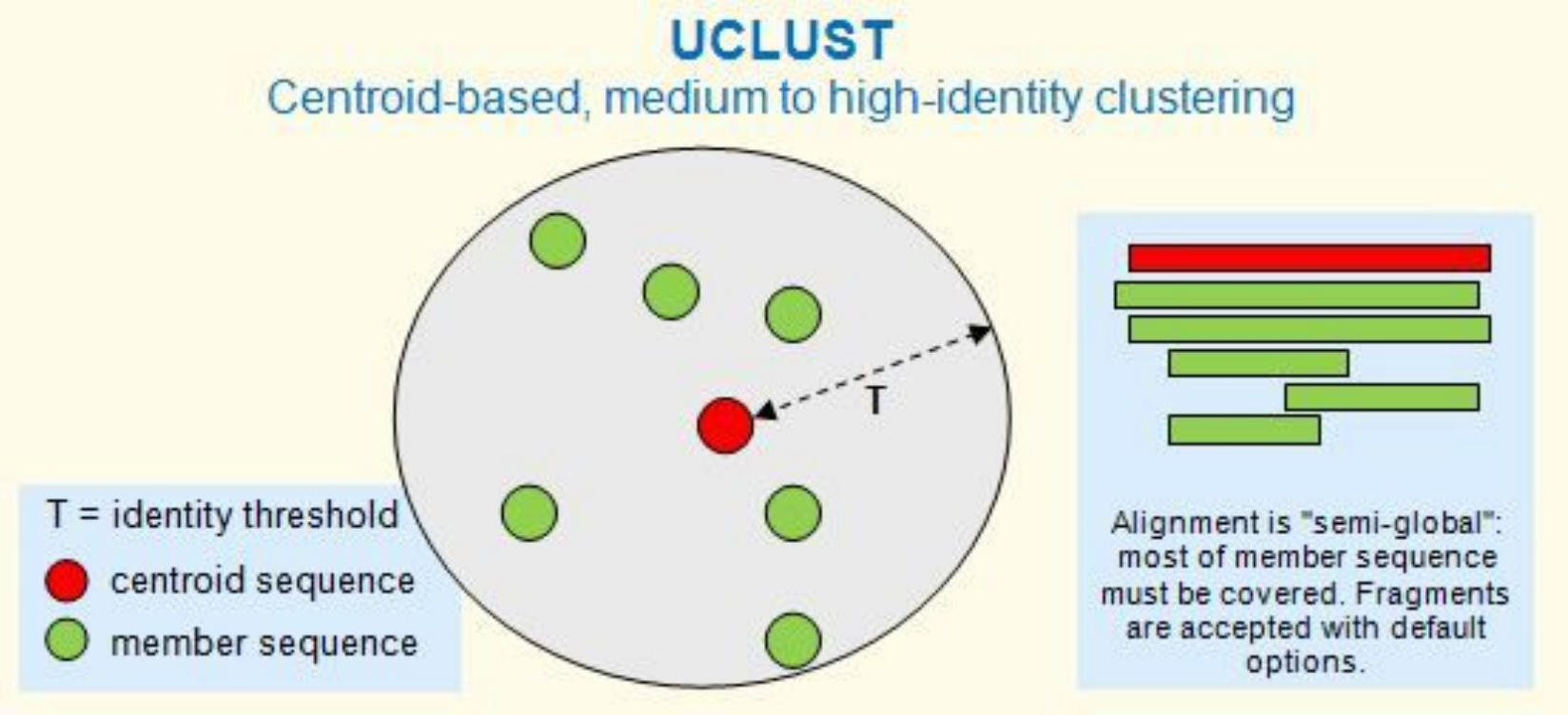


Jakou referenci použít pro KIRs?



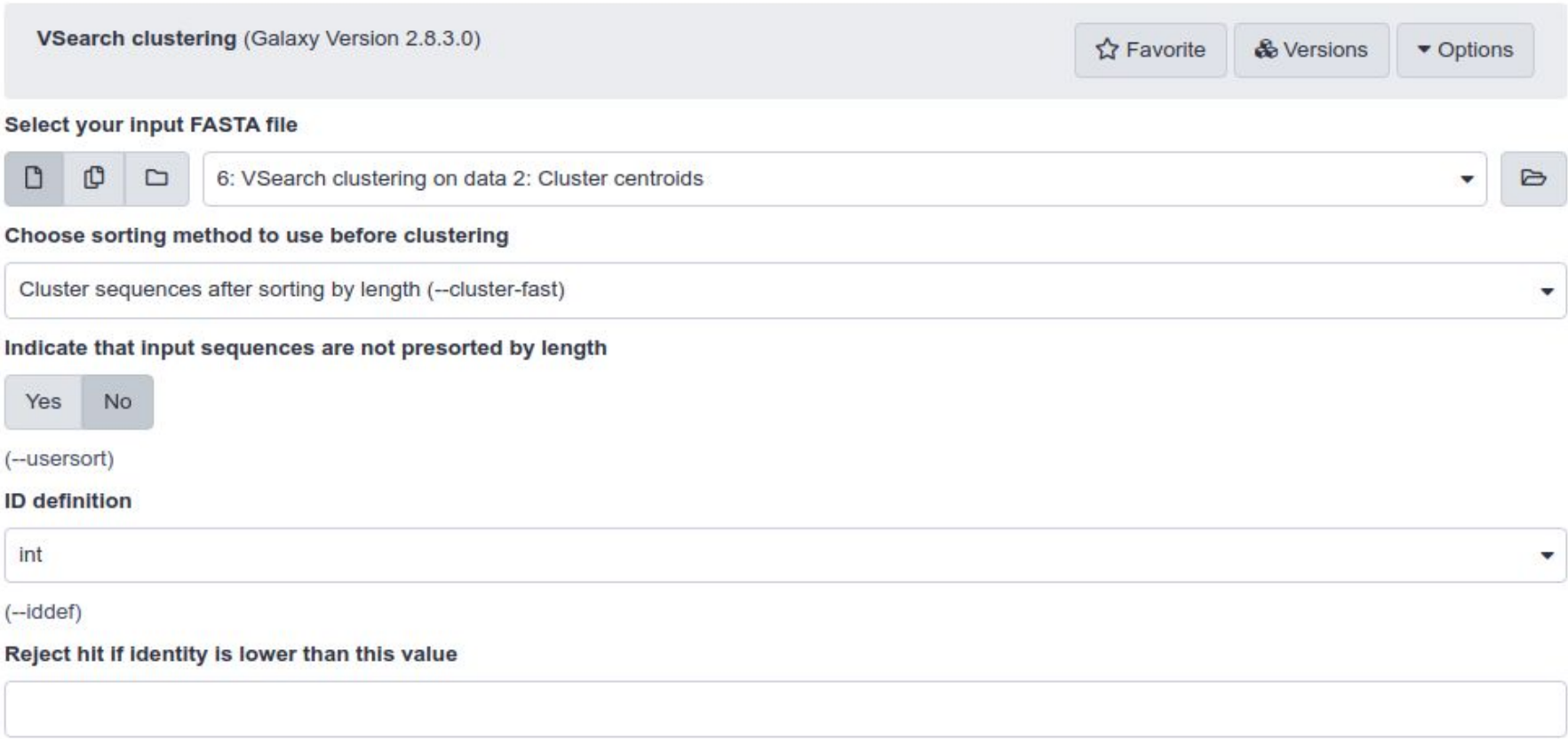
Jakou referenci použít pro KIRs?

Homology-based clustering



Jakou referenci použít pro KIRs?

Homology-based clustering



Cvičení_2 - clustering DNA sekvencí KIR receptorů

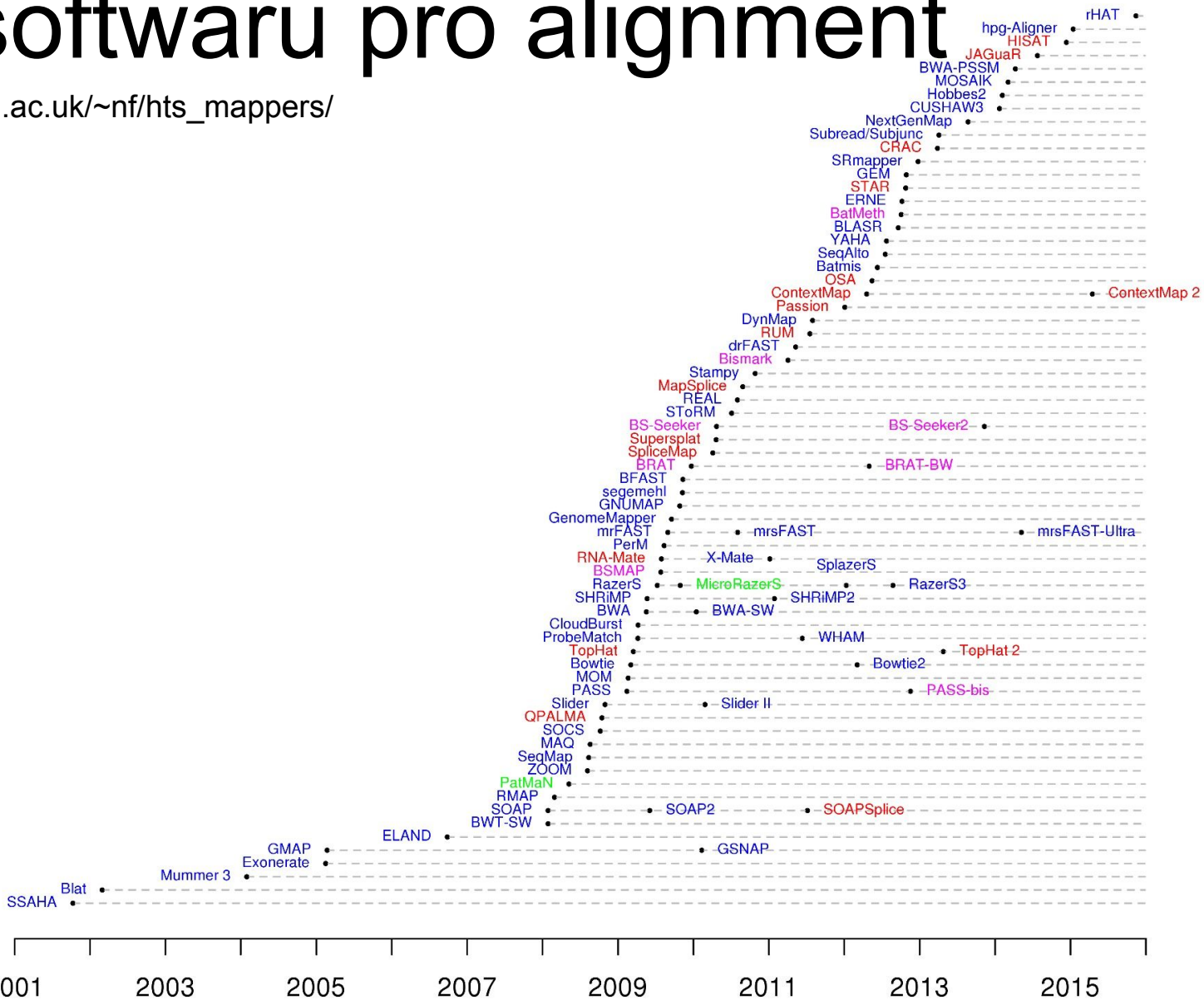
1. Importujte historii "***KIR-Clustering***" (https://usegalaxy.org/u/tomas_hron/h/kir-clustering).
2. Proveďte Clustering u obou souborů zvlášť pomocí ***"VSearch clustering"***:
   - Změňte verzi nástroje na Galaxy Version 1.9.7.0
   - Reject hit if identity is lower than this value: 0.95
   - Mask sequences: No masking
   - Write cluster abundances to centroid file. Yes
   - Select output files: Centroids

Kolik clusterů jste obdrželi? Jsou některé KIR geny rozděleny do více clusterů?

***https://usegalaxy.org/u/tomas_hron/h/kir-clustering-finished**

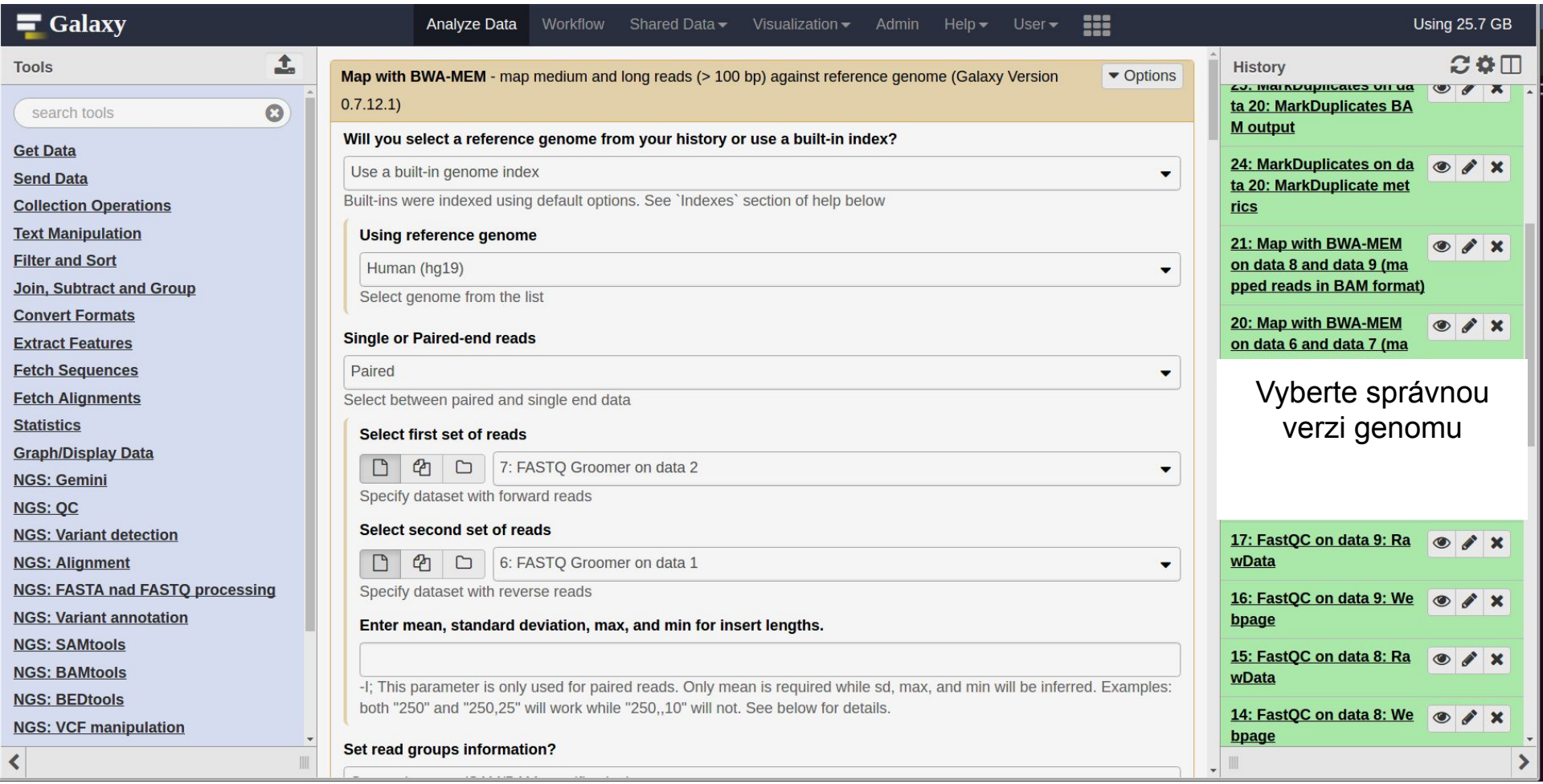
Vývoj softwaru pro alignment

http://www.ebi.ac.uk/~nf/hts_mappers/



BWA-MEM

- Je vylepšená verze Burrows-Wheeler Alignment algoritmu
- Optimalizován na čtení >= 70bp
- Je rychlejší než BWA a má lepší detekci variant, inserci a deleci
- Více informací na http://bio-bwa.sourceforge.net/


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 25.7 GB
Tools
search tools
Get Data
Send Data
Collection Operations
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data
NGS: Gemini
NGS: QC
NGS: Variant detection
NGS: Alignment
NGS: FASTA nad FASTQ processing
NGS: Variant annotation
NGS: SAMtools
NGS: BAMtools
NGS: BEDtools
NGS: VCF manipulation
Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.12.1)
Options
Will you select a reference genome from your history or use a built-in index?
Use a built-in genome index
Built-ins were indexed using default options. See `Indexes` section of help below
Using reference genome
Human (hg19)
Select genome from the list
Single or Paired-end reads
Paired
Select between paired and single end data
Select first set of reads
7: FASTQ Groomer on data 2
Specify dataset with forward reads
Select second set of reads
6: FASTQ Groomer on data 1
Specify dataset with reverse reads
Enter mean, standard deviation, max, and min for insert lengths.
-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.
Set read groups information?
History
ta 20: MarkDuplicates BA M output
24: MarkDuplicates on da ta 20: MarkDuplicate met rics
21: Map with BWA-MEM on data 8 and data 9 (ma pped reads in BAM format)
20: Map with BWA-MEM on data 6 and data 7 (ma
Vyberte správnou verzi genomu
17: FastQC on data 9: Ra wData
16: FastQC on data 9: We bpage
15: FastQC on data 8: Ra wData
14: FastQC on data 8: We bpage

Cvičení 3 - mapování DNA knihovny KIR receptorů

1. Importujte historii "***KIR-mapping***" (https://usegalaxy.org/u/tomas_hron/h/kir-mapping).
2. Proveďte 3 různá mapování pomocí nástroje "***Map with BWA-MEM***" s referenční sekvencí:

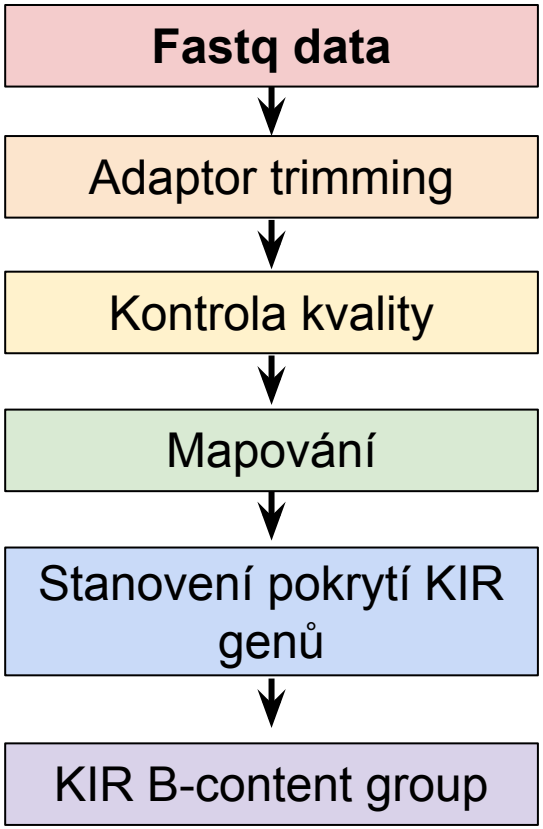
   i) KIR_gen-clustered
   ii) KIR_nuc-clustered
   iii) Human: hg38 Canonical

3. Pro každý BAM soubor spusťte nástroj "***Samtools flagstat***" a zjistěte kolik procent readů úspěšně zamapovalo

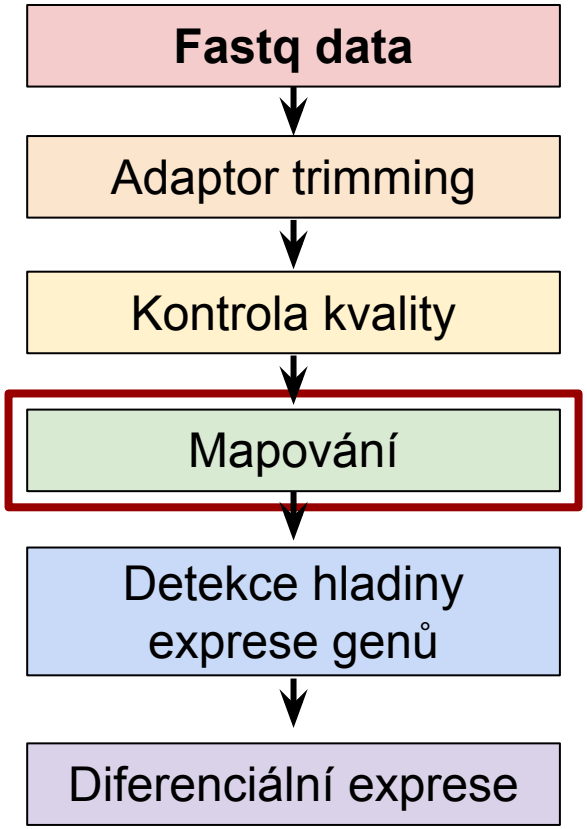
   Která reference se zdá být nejvhodnější?

***https://usegalaxy.org/u/tomas_hron/h/kir-mapping-finished**

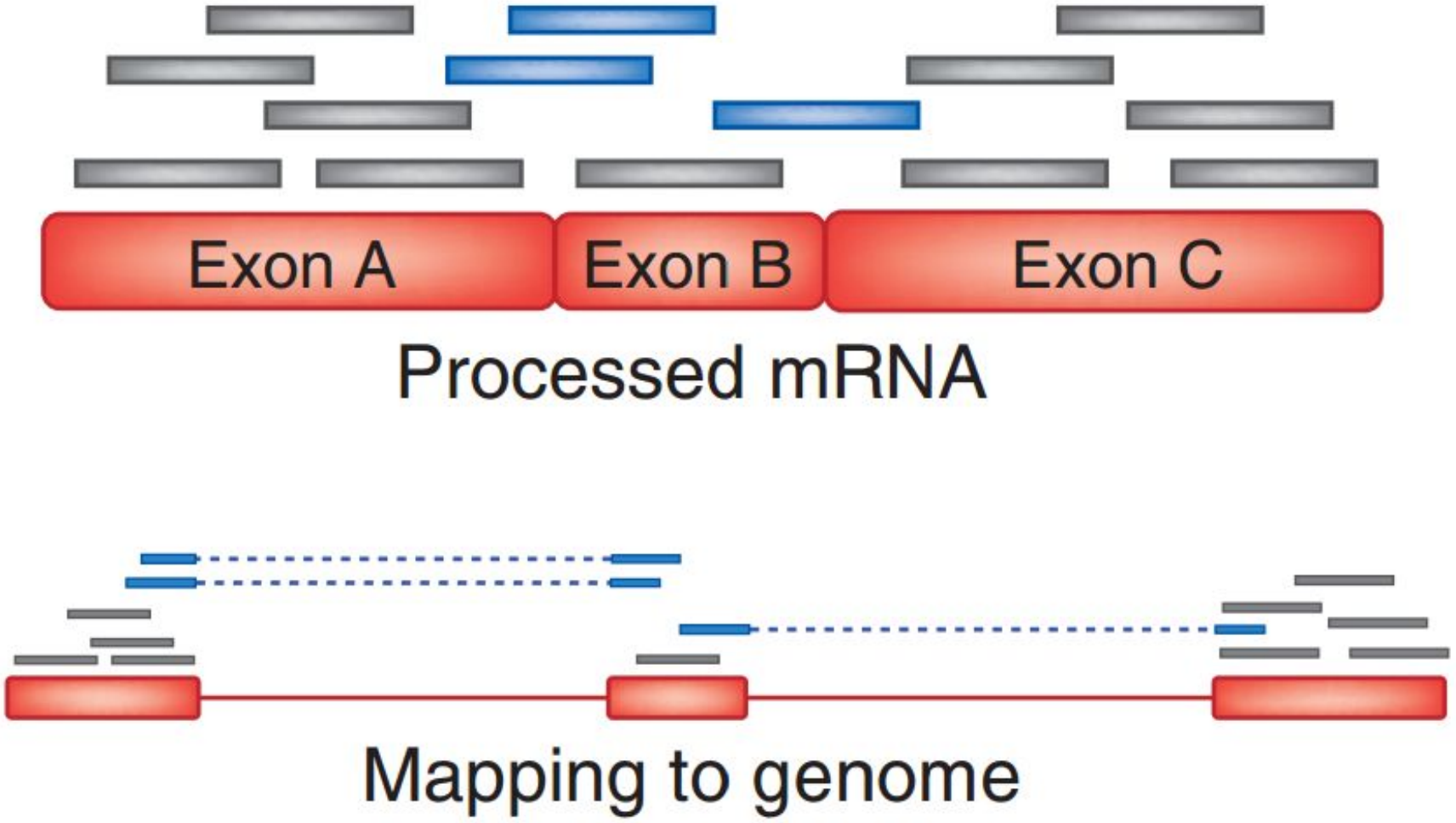
Genotypizace KIRs



Diferenciální exprese genů u D. melanogaster

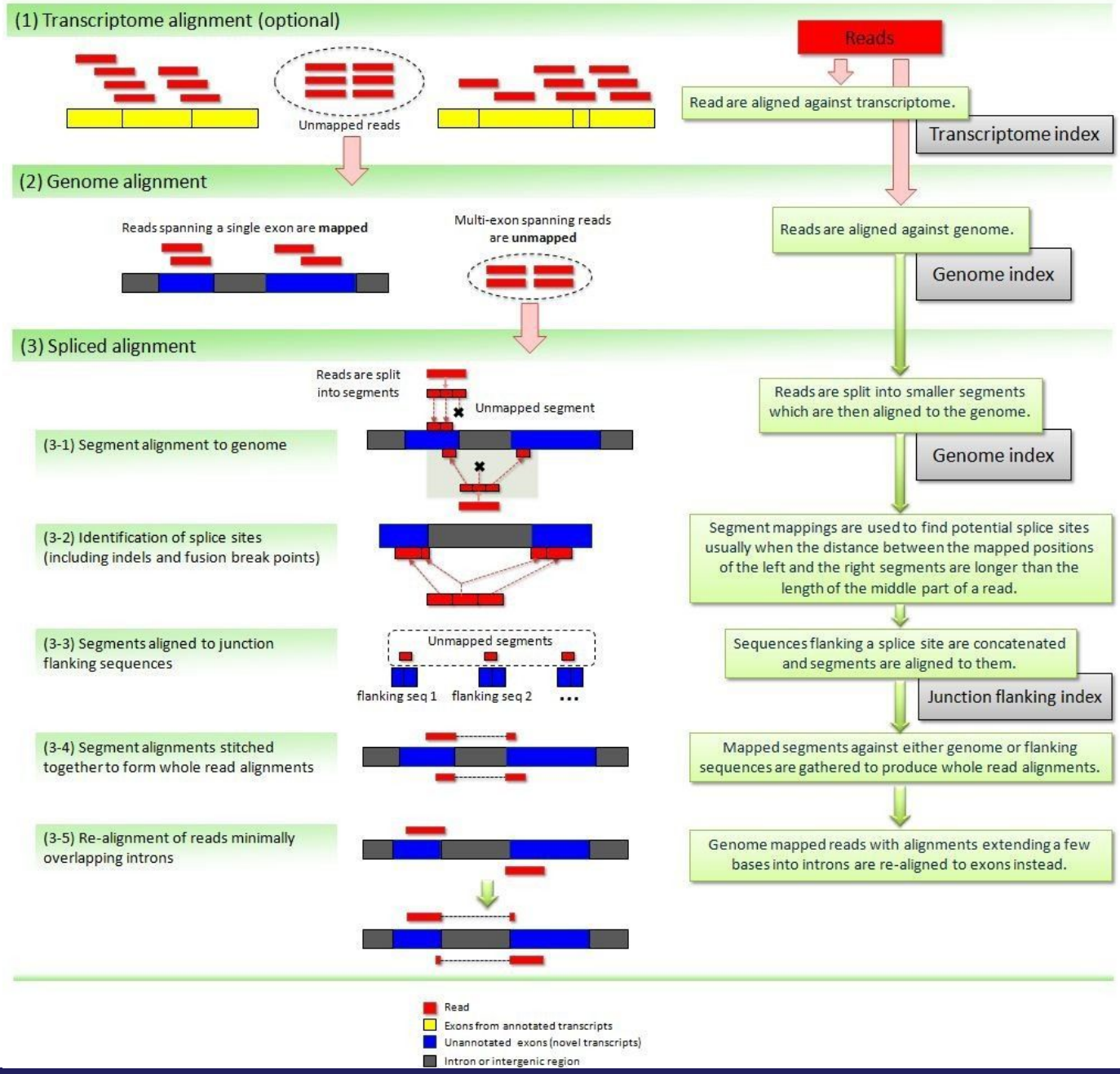


RNA mapping

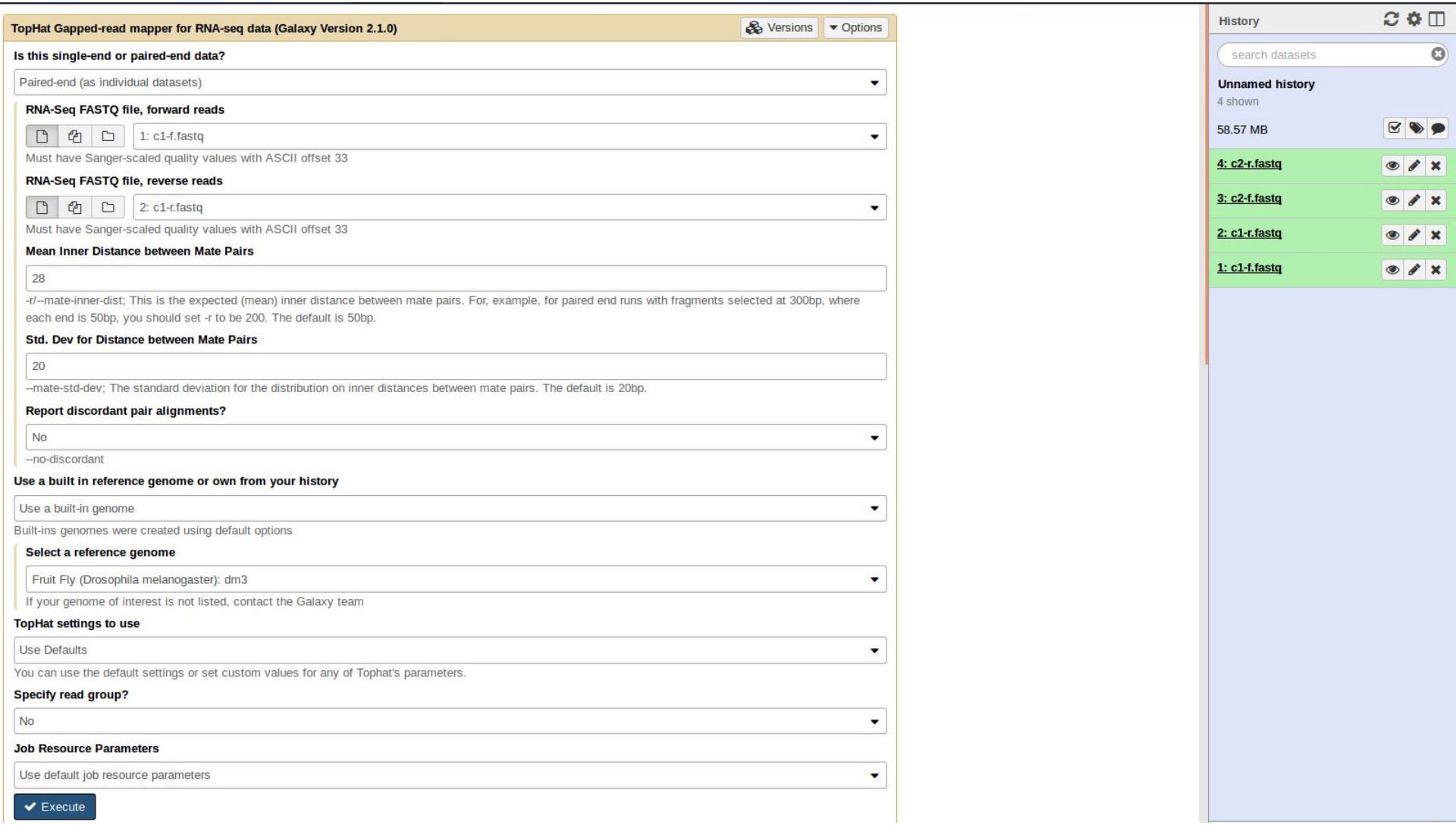


Trapnell a Salzberg, 2009

TopHat2


(1) Transcriptome alignment (optional)
Unmapped reads
Reads
Read are aligned against transcriptome.
Transcriptome index
(2) Genome alignment
Reads spanning a single exon are mapped
Multi-exon spanning reads are unmapped
Reads are aligned against genome.
Genome index
(3) Spliced alignment
Reads are split into segments
Unmapped segment
Reads are split into smaller segments which are then aligned to the genome.
Genome index
(3-1) Segment alignment to genome
(3-2) Identification of splice sites (including indels and fusion break points)
Segment mappings are used to find potential splice sites usually when the distance between the mapped positions of the left and the right segments are longer than the length of the middle part of a read.
(3-3) Segments aligned to junction flanking sequences
Unmapped segments
flanking seq 1
flanking seq 2
...
Sequences flanking a splice site are concatenated and segments are aligned to them.
Junction flanking index
(3-4) Segment alignments stitched together to form whole read alignments
Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.
(3-5) Re-alignment of reads minimally overlapping introns
Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.
Read
Exons from annotated transcripts
Unannotated exons (novel transcripts)
Intron or intergenic region

TopHat

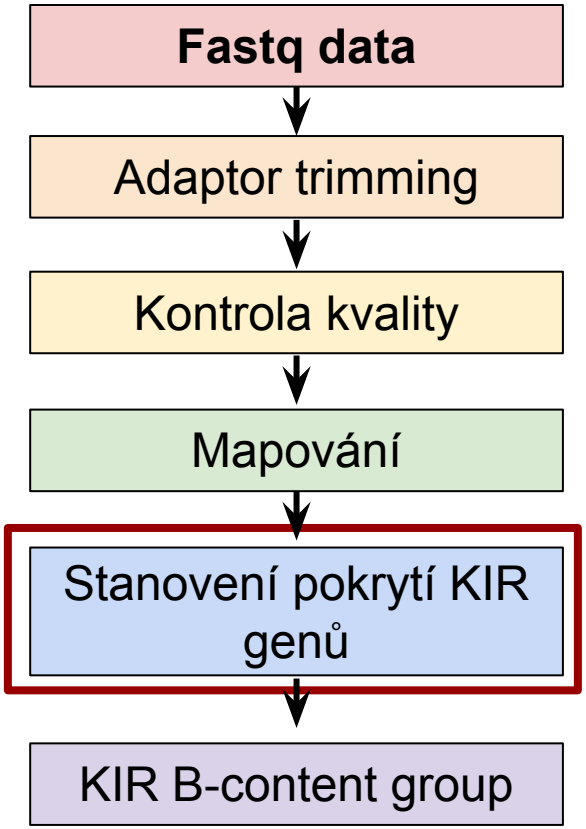

TopHat Gapped-read mapper for RNA-seq data (Galaxy Version 2.1.0)
Versions
Options
Is this single-end or paired-end data?
Paired-end (as individual datasets)
RNA-Seq FASTQ file, forward reads
1: c1-f.fastq
Must have Sanger-scaled quality values with ASCII offset 33
RNA-Seq FASTQ file, reverse reads
2: c1-r.fastq
Must have Sanger-scaled quality values with ASCII offset 33
Mean Inner Distance between Mate Pairs
28
-r/--mate-inner-dist; This is the expected (mean) inner distance between mate pairs. For, example, for paired end runs with fragments selected at 300bp, where each end is 50bp, you should set -r to be 200. The default is 50bp.
Std. Dev for Distance between Mate Pairs
20
--mate-std-dev; The standard deviation for the distribution on inner distances between mate pairs. The default is 20bp.
Report discordant pair alignments?
No
--no-discordant
Use a built in reference genome or own from your history
Use a built-in genome
Built-ins genomes were created using default options
Select a reference genome
Fruit Fly (Drosophila melanogaster): dm3
If your genome of interest is not listed, contact the Galaxy team
TopHat settings to use
Use Defaults
You can use the default settings or set custom values for any of Tophat's parameters.
Specify read group?
No
Job Resource Parameters
Use default job resource parameters
Execute
History
search datasets
Unnamed history
4 shown
58.57 MB
4: c2-r.fastq
3: c2-f.fastq
2: c1-r.fastq
1: c1-f.fastq

Cvičení_4 - mapování RNAseq

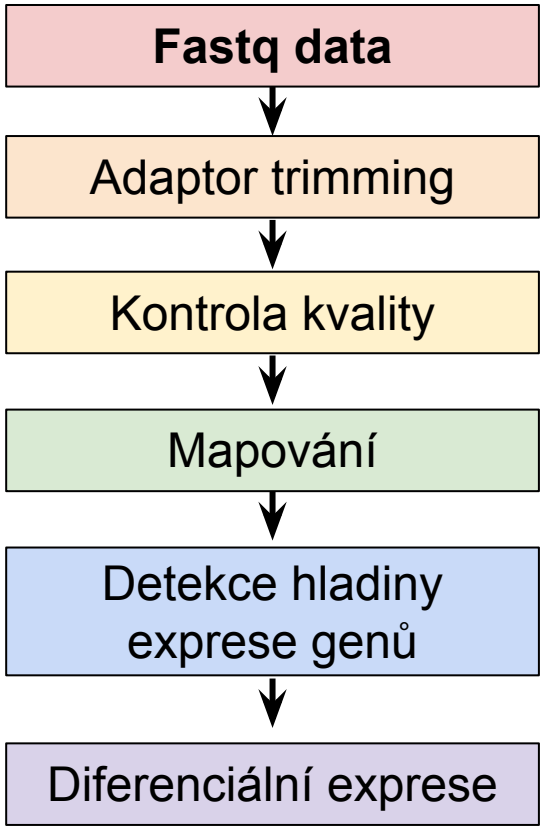
1. Importujte historii "***RNA-Mapping***" (https://usegalaxy.org/u/tomas_hron/h/rna-mapping).
2. Proveďte mapování pomocí programu ***"TopHat"***, nastavte parametry:
   - RNA-Seq FASTQ file, forward reads: c1-r1-f-x.fastq
   - RNA-Seq FASTQ file, reverse reads: c1-r1-r-x.fastq
   - Mean Inner Distance between Mate Pairs = 28
   - Use built-in genome
   - Select the reference genome = dm6
   - TopHat settings to use = Full parameter list (This is done to be able to specify the strandedness of the library)
   - Library Type = FR First Strand
3. **Odvodíte, co obsahují výsledné soubory?**
4. Otevřete BAM soubor s příponou "***accepted hits***" v IGV (https://igv.org/app/) a zobrazte oblast ***chrX:11,801,434-11,807,435***. **Dokážete najít čtení, která mapují přes dva introny (split reads)?**
5. Klikněte pravým tlačítkem do prostoru alignmentu v IGV a vyberte možnost ***Sashimi Plot***. **Domyslíte, co zobrazený graf ukazuje?**

*https://usegalaxy.org/u/tomas_hron/h/rna-mapping-finished

Genotypizace KIRs



Diferenciální exprese genů u D. melanogaster



SAM FLAGs

#	Binary	Decimal	Hexadecimal	Description
1	1	1	0x1	Read paired
2	10	2	0x2	Read mapped in proper pair
3	100	4	0x4	Read unmapped
4	1000	8	0x8	Mate unmapped
5	10000	16	0x10	Read reverse strand
6	100000	32	0x20	Mate reverse strand
7	1000000	64	0x40	First in pair
8	10000000	128	0x80	Second in pair
9	100000000	256	0x100	Not primary alignment
10	1000000000	512	0x200	Read fails platform/vendor quality checks
11	10000000000	1024	0x400	Read is PCR or optical duplicate
12	100000000000	2048	0x800	Supplementary alignment

http://www.samformat.info/sam-format-flag

SAM FLAGs

Kombinace bitových FLAGů

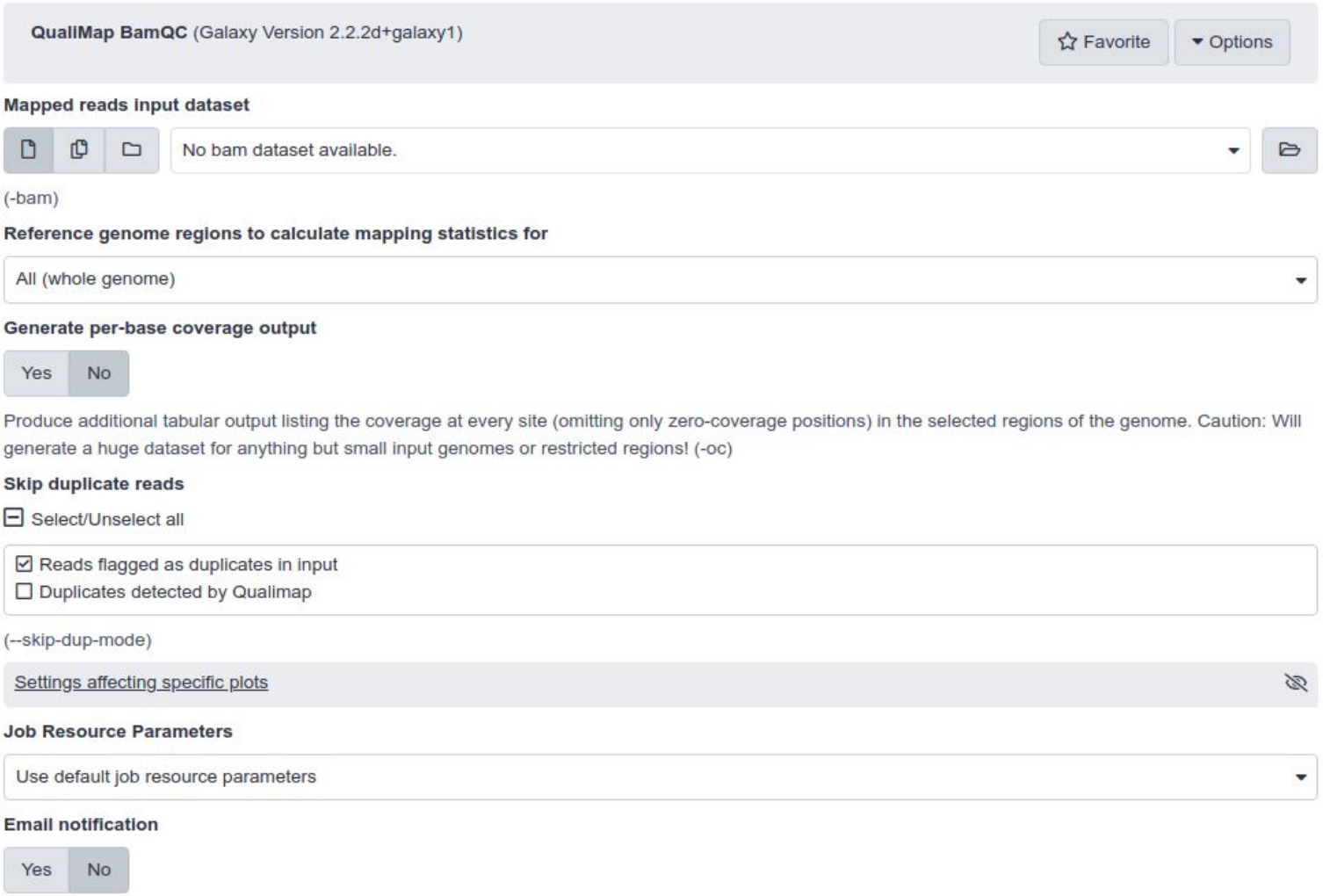
Příklad: FLAG = 65 čtení je první v páru

párová čtení ($2^0 = 1$) + první v páru ($2^8 = 64$) = 65

Vyzkoušejme si to zde:
http://www.samformat.info/sam-format-flag

QualiMap BamQC

statistika mapování



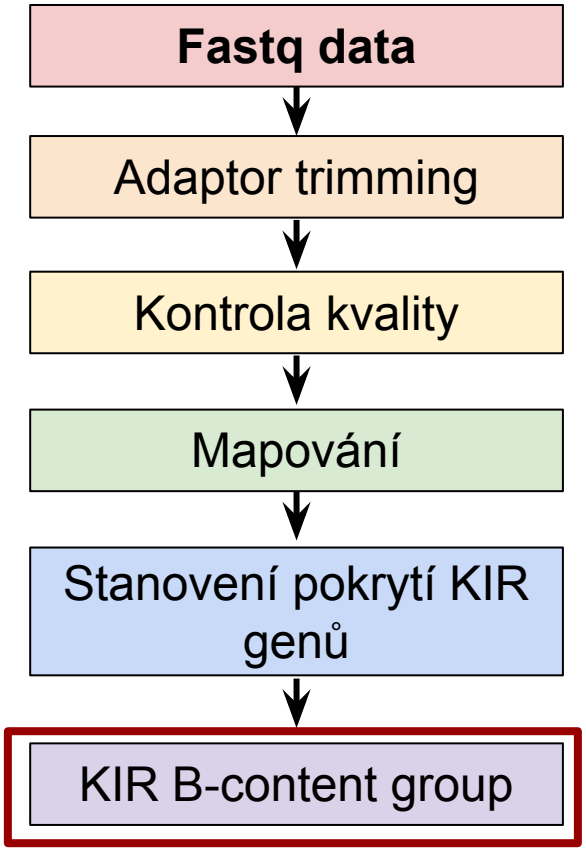
Cvičení 5 - Statistika mapování a Filtrování BAM souboru

1. Importujte historii "**KIR-Genotyping**" (https://usegalaxy.org/u/tomas_hron/h/kir-genotyping).
2. Spusťte nástroj ***Filter BAM datasets*** a vyberte BAM soubor, v kterém jsou použity jako reference clusterované genové sekvence KIR. Odfiltrujte všechny ready, které nemapují k referenci unikátně (Mapping Quality = 0)
   - pro filtrování použijte hodnotu MapQ >=1
3. Pro původní i výsledný BAM soubor použijte nástroj ***QualiMap BamQC***

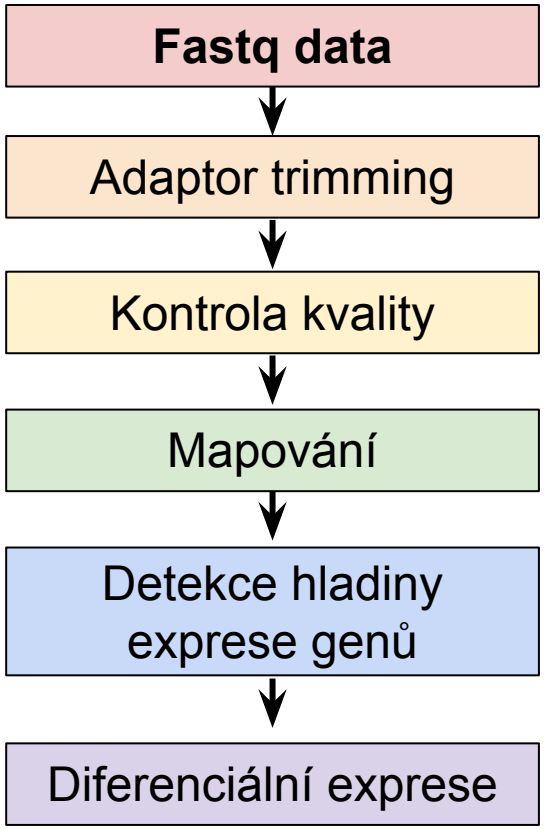
Kolik readů nám zbylo?
Které KIR geny jsou sekvenačně pokryty unikátními ready?

***https://usegalaxy.org/u/tomas_hron/h/kir-genotyping-finished**

Genotypizace KIRs



Diferenciální exprese genů u D. melanogaster



Cvičení 6 - Donor KIR B-content group calculator

Donor KIR B-content group calculator

Killer-cell Immunoglobulin-like Receptor (KIR) genes form a diverse, immunogenetic system unlinked to HLA. Group A and B KIR haplotypes have distinctive centromeric (Cen) and telomeric (Tel) gene-content motifs. With the goal of developing a donor selection strategy to improve transplant outcome, Cooley *et al.*, compared the contribution of these motifs to the clinical benefit conferred by B haplotype donors. Donor KIR genotype influenced transplantation outcome for AML, but not ALL, after HLA-matched or mismatched T-cell replete unrelated donor transplants. Compared to A haplotype motifs, centromeric and telomeric B motifs both contributed to relapse protection and improved survival, but Cen-B homozygosity had the strongest independent effect. KIR genotyping several best HLA-matched potential donors should substantially increase the frequency of transplants using unrelated donor grafts with favorable KIR gene content. Adopting this practice could result in superior disease-free survival for patients transplanted for AML.

Disclaimer - This research tool is being offered as a tool to predict donor KIR B-content groups assignments as reported in:

S Cooley, DJ Weisdorf, LA Guethlein, JP Klein, T Wang, CT Le, SGE Marsh, D Geraghty, S Spellman, MD Haagenson, M Ladner, E Trachtenberg, P Parham and JS Miller.
Donor selection for natural killer cell receptor genes leads to superior survival after unrelated transplantation for acute myelogenous leukemia.
Blood (2010) 116:2411-9.

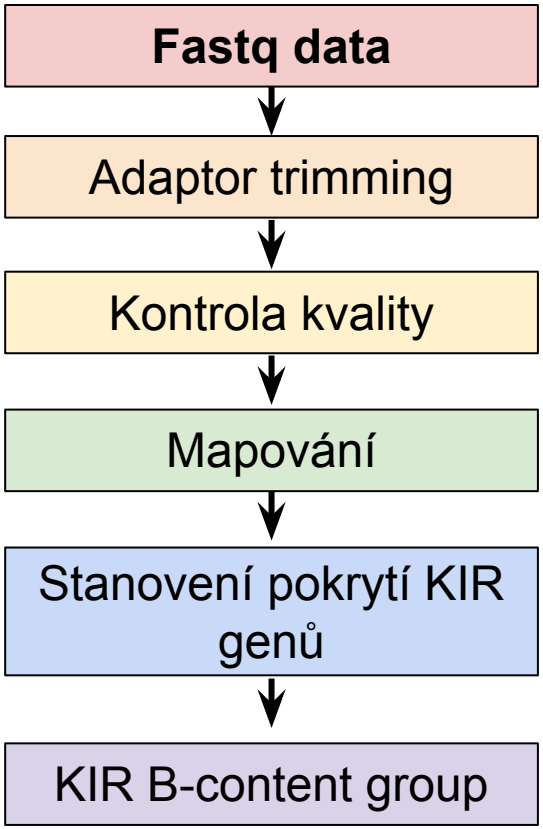
No information entered into this tool is collected or stored on our servers.

This calculator allows you to enter the genotypes for up to five prospective donors, and receive their assignments to one of 3 groups based on KIR B-content. The groups, "Neutral", "Better", "Best", refer to the associated relapse protection seen in T-cell replete URD HCT for AML. Simply enter the prospective donor's KIR genotype by selecting the appropriate box for each gene. A tick indicates presence of a gene in the donor's genotype.

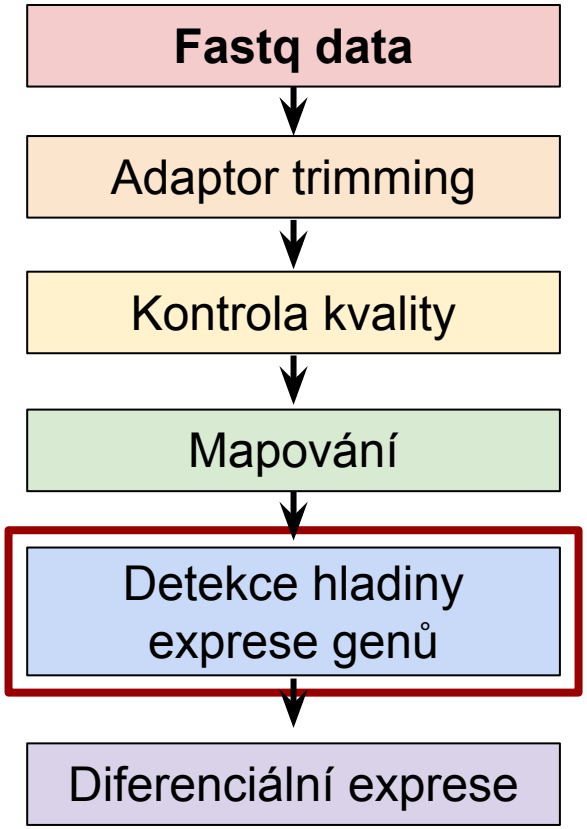
Resources

About > | Statistics >
Publications > | Nomenclature >
Genes > | Alleles >
Haplotypes > | Genotypes >
Alignments > | Releases >
BLAST > | B-Content >
Ligands > | FTP >
GitRepos > | Cells >
FAQ > | Submissions >

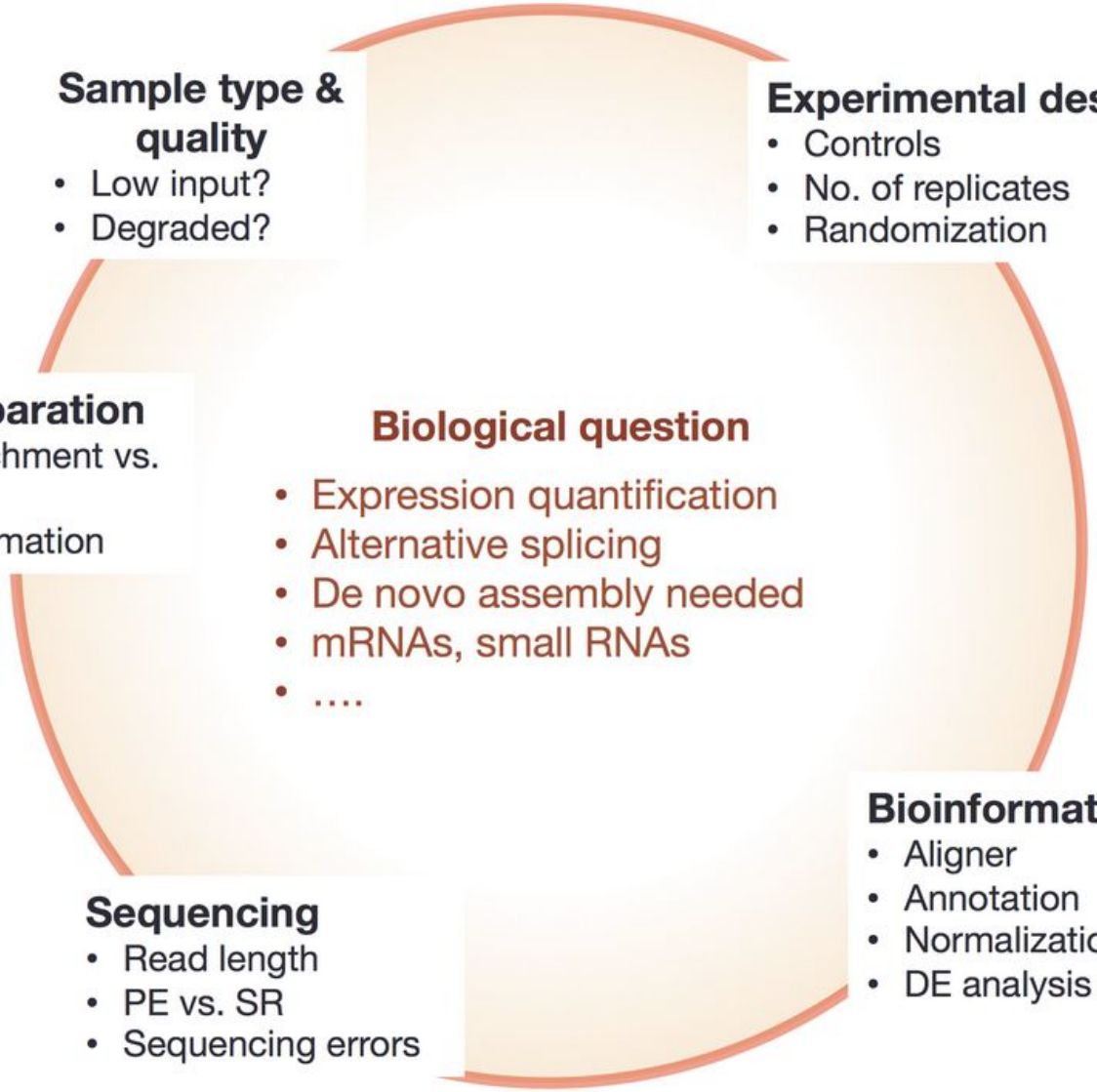
Genotypizace KIRs

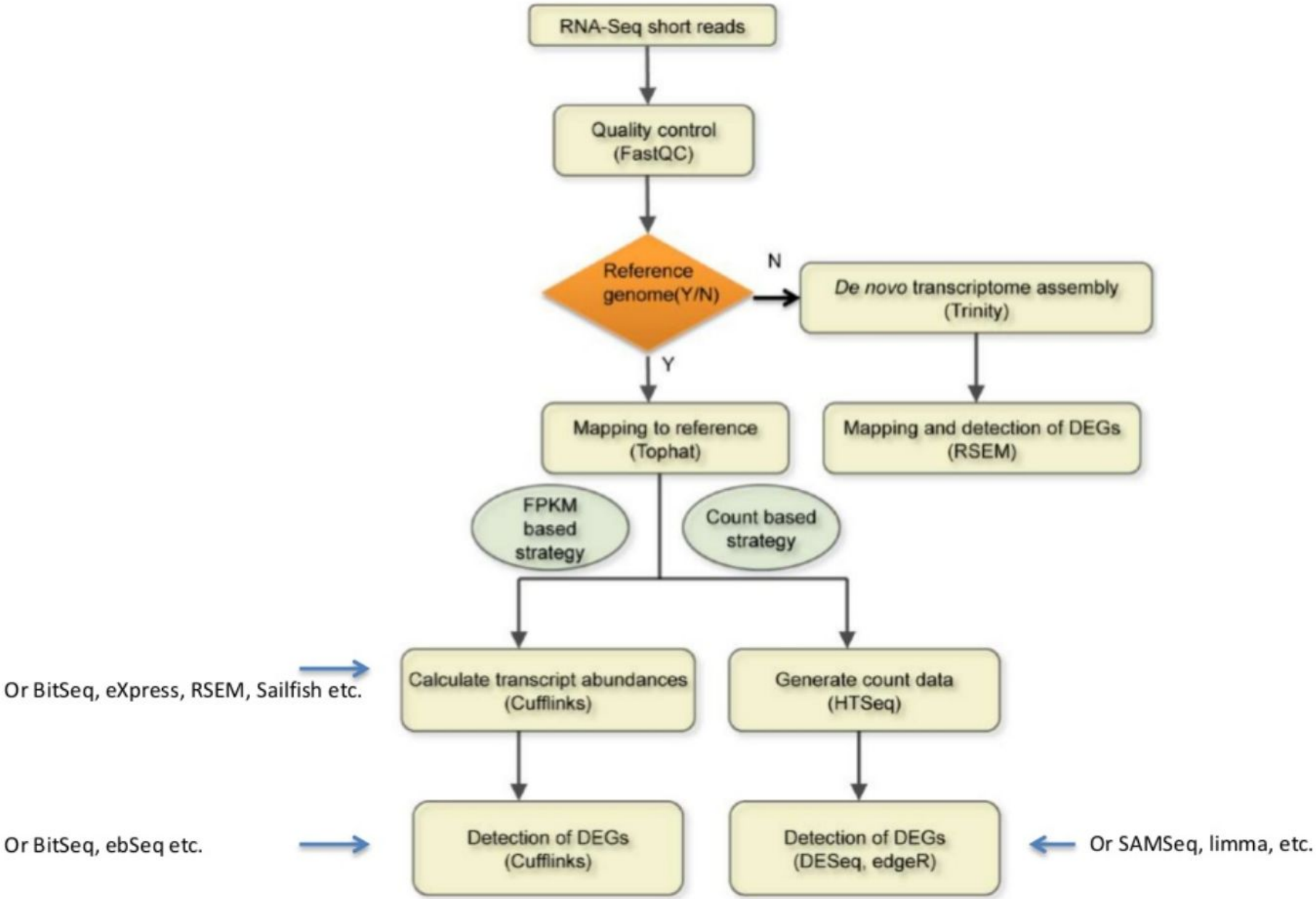


Diferenciální exprese genů u D. melanogaster



Everything's connected…

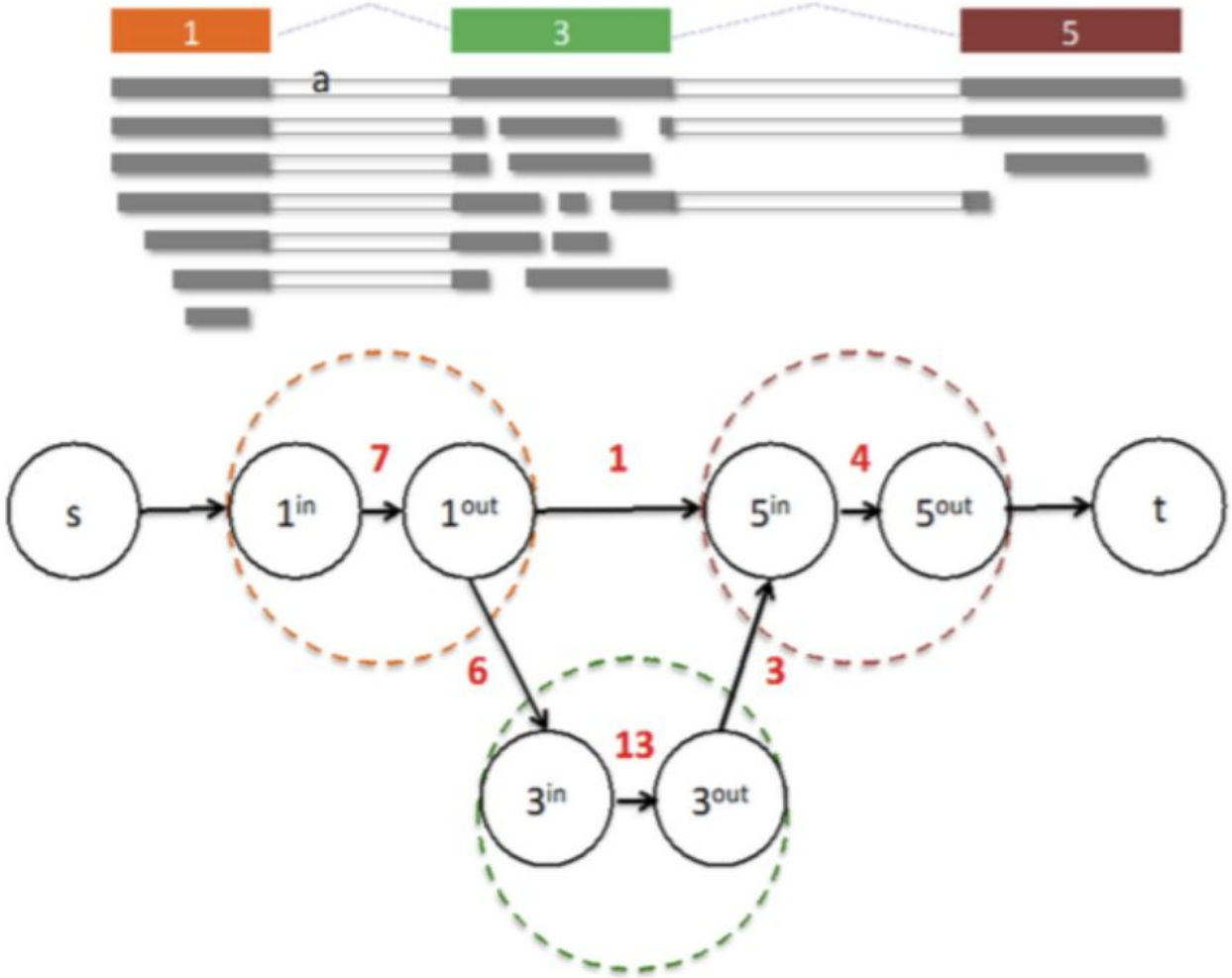



RNA-Seq short reads
Quality control (FastQC)
Reference genome(Y/N)
N
De novo transcriptome assembly (Trinity)
Y
Mapping to reference (Tophat)
Mapping and detection of DEGs (RSEM)
FPKM based strategy
Count based strategy
Or BitSeq, eXpress, RSEM, Sailfish etc.
Calculate transcript abundances (Cufflinks)
Generate count data (HTSeq)
Or BitSeq, ebSeq etc.
Detection of DEGs (Cufflinks)
Detection of DEGs (DESeq, edgeR)
Or SAMSeq, limma, etc.

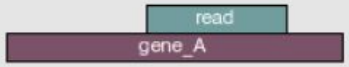
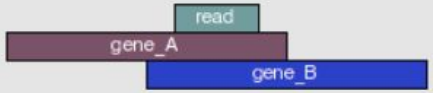
From reads to RNA molecules

- **Reference genome-based** - an assembled genome exists for a species for which an RNAseq experiment is performed. It allows reads to be aligned against the reference genome and significantly improves our ability to reconstruct transcripts. This category would obviously include humans and most model organisms.
- **Reference genome-free** - no genome assembly for the species of interest is available. In this case one would need to assemble the reads into transcripts using *de novo* approaches.

Transcript counting (StringTie)



Gene counting (HTseq-count)

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Cvičení 7 - Stanovení genové exprese

1. Importujte historii “RNA-DEA” (https://usegalaxy.org/u/tomas_hron/h/rna-dea), kde naleznete 6 BAM souborů (dvě experimentální podmínky v triplikátu)

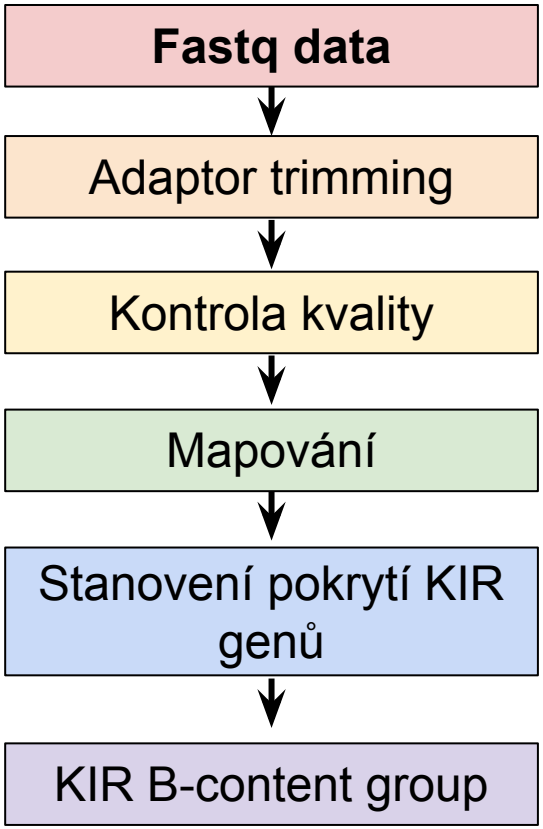
   Condition1_replicate1, Condition1_replicate2, Condition1_replicate3

   Condition2_replicate1, Condition2_replicate2, Condition2_replicate3

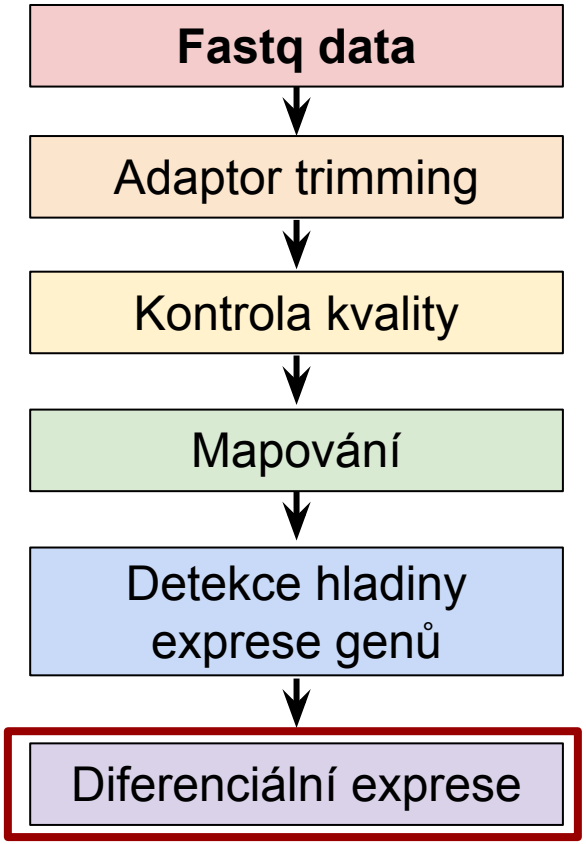
2. Vytvořte tabulku počtu čtení pro jednotlivé geny pomocí nástroje “***htseq-count***”. Vyberte všechny BAM soubory, zvolte “model=UNION”, “Strand=yes”, “Minimal align quality=10”.

*https://usegalaxy.org/u/tomas_hron/h/rna-dea-finished

Genotypizace KIRs



Diferenciální exprese genů u D. melanogaster



Read count normalisation (within sample)

Table 11: Normalization methods for the comparison of gene read counts within the same sample.

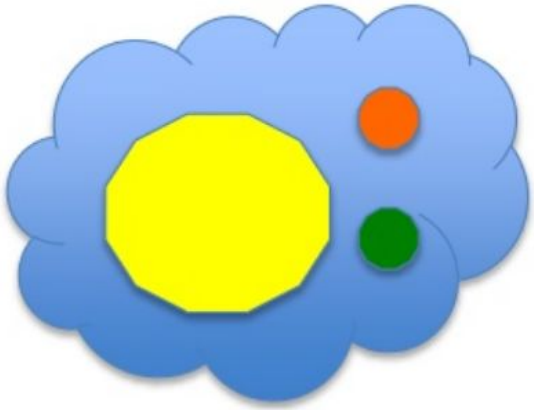
Name	Details	Comment
RPKM (reads per kilobase of exons per million mapped reads)	1. For each gene, count the number of reads mapping to it. 2. Divide that count by: the length of the gene in base pairs divided by 1,000 multiplied by the total number of mapped reads divided by 10^6. $$RPKM_i = \frac{read\,count\,of\,gene\,i}{(\frac{length\,of\,gene\,i}{10^3})(\frac{library\,size}{10^6})}$$	• introduces a bias in the per-gene variances, in particular for lowly expressed genes (Oshlack and Wakefield, 2009) • implemented in `edgeR`'s `rpkm()` function
FPKM (fragments per kilobase...)	1. Same as RPKM, but for paired-end reads: 2. The number of fragments (defined by two reads each) is used.	• implemented in `DESeq2`'s `fpkm()` function
TPM	Instead of normalizing to the total library size, TPM represents the abundance of an individual gene i in relation to the abundances of the other transcripts (e.g., j) in the sample. 1. For each gene, count the number of reads mapping to it and divide by its length in base pairs (= counts per base). 2. Multiply that value by 1 divided by the sum of all counts per base of every gene. 3. Multiply that number by 10^6. $$TPM_i = \frac{X_i}{l_i} * \frac{1}{\sum_j \frac{X_j}{l_k}}$$	• details in Wagner et al. (2012)

TMM – Trimmed Mean of M values

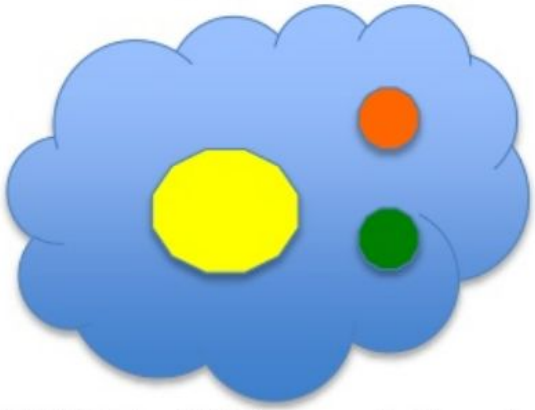
Attempts to correct for differences in RNA *composition* between samples

E g if certain genes are very highly expressed in one tissue but not another, there will be less "sequencing real estate" left for the less expressed genes in that tissue and RPKM normalization (or similar) will give biased expression values for them compared to the other sample

RNA population 1



RNA population 2



Equal sequencing depth -> orange and red will get lower RPKM in RNA population 1 although the expression levels are actually the same in populations 1 and 2

Robinson and Oshlack Genome Biology 2010, 11:R25, http://genomebiology.com/2010/11/3/R25

Differential expression analysis - DEseq

DESeq2 Determines differentially expressed features from count tables (Galaxy Version 2.1.8.3) — Options

Factor

1: Factor

Specify a factor name

Conditions

Only letters, numbers and underscores will be retained in this field

Factor level

1: Factor level

Specify a factor level

Condition 1

Only letters, numbers and underscores will be retained in this field

Counts file(s)

84: htseq-count on collection 37

2: Factor level

Specify a factor level

Condition 2

Only letters, numbers and underscores will be retained in this field

Counts file(s)

92: htseq-count on collection 57

+ Insert Factor level

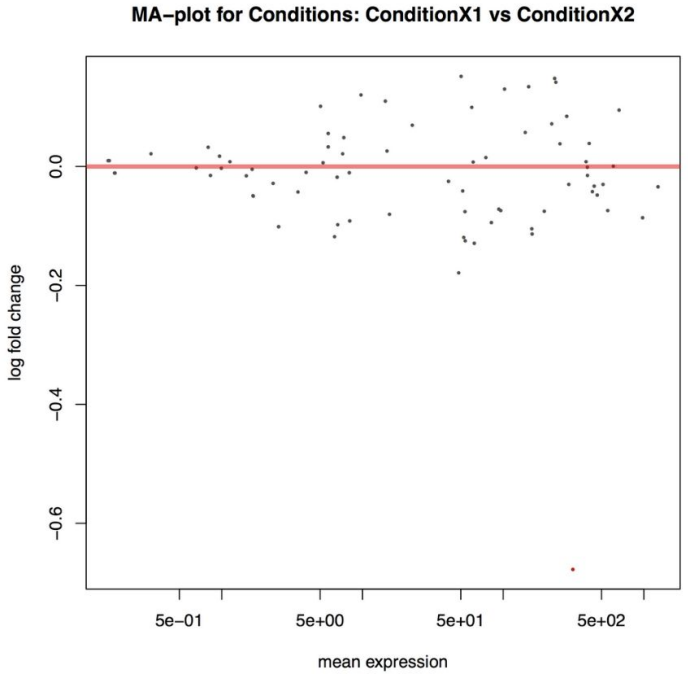
+ Insert Factor

Visualising the analysis results

Yes | No

This will produce output as shown below. The columns are: (**1**) gene identifier, (**2**) mean normalised counts, averaged over all samples from both conditions, (**3**) logarithm (base 2) of the fold change, (**4**) the standard error estimate for the log2 fold change estimate, (**5**) Wald test statistic, (**6**) p value for the statistical significance of this change, and (**7**) *p*-value adjusted for multiple testing with the Benjamini-Hochberg procedure which controls false discovery rate (FDR). There is only one gene with significant change in gene expression between conditions: `CG1803-RC` with *p*-value = 1.6×10^{-05}

1	2	3	4	5	6	7
CG1803-RC	313.399854993883	-0.677347951975255	0.130547283480314	-5.18852582694606	2.11965367545954e-07	1.61093679334925e-05
CG10352-RA	3.97546094727023	-0.00977069078261261	0.0747042077056103	-0.130791706152836	0.895940084165715	0.981157785658315
CG10353-RD	22.6113188342966	0.0696827956291263	0.129706150655657	0.537235861806736	0.591104702591145	0.981157785658315
CG10362-RA	50.2265915083909	0.151864019842212	0.137485127753392	1.10458507275501	0.269339422115096	0.981157785658315



Cvičení 8 - Analýza diferenciální exprese

1. Otevřete historii “RNA-DEA” (https://usegalaxy.org/u/tomas_hron/h/rna-dea).
2. Proveďte analýzu diferenciální exprese pomocí “***DESeq2***” (změňte na verzi 2.1.8.3). Data obsahují dva vzorky ve třech replikátech. Nastavte název faktoru = “Condition” a vytvořte dvě úrovně faktoru (Factor level) - Condition1 a Condition2. Do těchto úrovní vyplňte příslušná data. “Visualising analysis results=yes”, “Output all levels vs all levels=no”.
3. Ve výsledné tabulce jsou hodnoty pro jednotlivé proměnné: (**1**) gene identifier, (**2**) mean normalised counts, averaged over all samples, (**3**) log2 of the fold change, (**4**) the standard error estimate for the log2 fold change, (**5**) Wald test statistic, (**6**) p value for the statistical significance of this change, (**7**) *p*-value adjusted for multiple testing with the Benjamini-Hochberg procedure which controls false discovery rate (FDR).
4. **Je nějaký gen ve vzorcích diferenciálně exprimován (signifikantně)?**

RNAseq analysis tutorial

RNAseq: Reference-based

This tutorial is inspired by an exceptional RNAseq course at the Weill Cornell Medical College compiled by Friederike Dündar, Luce Skrabanek, and Paul Zumbo and by tutorials produced by Björn Grüning (@bgruening) for Freiburg Galaxy instance. Much of Galaxy-related features described in this section have been developed by Björn Grüning (@bgruening) and configured by Dave Bouvier (@davebx).

- https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/rb-rnaseq/tutorial.html

- DATA:

  usegalaxy.org:
  Shared_Data/Data_Libraries/Tutorials/RNA_seq_(RB)

www.hpst.cz

Autorizovaný distributor Agilent Technologies

Agilent Authorized Distributor

Produkty
Služby

Kontakty
Servisní požadavek

O nás
Kariéra

Knihovna Agilent
Odkazy

Novinky

Semináře

21. 3. 2018 Novinky a trendy Agilent Technologies 2018

Workshop

23.1.-24.1. 2018 Analýza dat z NGS s Galaxy bez znalosti příkazového řádku

Nové produkty

3D tomografický fluorescenční mikroskop od Nanolive

Nové aplikace

ICP-QQQ - komplexní online knihovna aplikací

Školení

MassHunter pro GC/MS

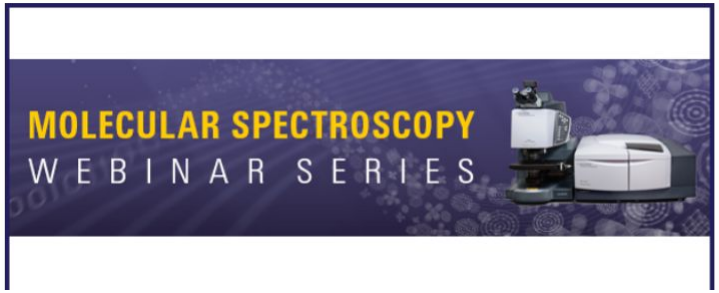
OpenLAB CDS 2.x

LC školení, uživatelské dovednosti a diagnostika

GC školení, uživatelské dovednosti a diagnostika

NOVINKA: GC školení, uživatelské dovednosti a diagnostika pro pokročilé uživatele

MS (single-quad/QQQ pro GC) školení, uživatelské dovednosti a diagnostika



thron@hpst.cz
dgg@hpst.cz